Seeking Truly Novel Proteins in a Fully Random Library by Bacterial Tat-dependent Selections for Folding

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von
Mark Alexander Schmitz
aus
Deutschland

Promotionskomitee
Prof. Dr. Andreas Plückthun (Vorsitz)
Prof. Dr. Ben Schuler
Prof. Dr. Elke Deuerling

Zürich, 2015

Erklärung

Diese Dissertation wurde selbstständig, ohne unerlaubte Hilfe im Sinne von §3 und §5 der Promotionsverordnung vom 08. Juli 2002, angefertigt. Bei der Abfassung der Dissertation wurden keine anderen als die darin angegebenen Hilfsmittel benutzt.

Zürich, Oktober 2014

Mark A. Schmitz

Zusammenfassung:

Proteine sind unersetzbare Bestandteile aller uns bekannten Lebewesen. Die meisten biologischen Prozesse könnten nicht ohne diese faszinierenden Makromoleküle ablaufen. Ein Protein besteht aus Aminosäuren, und die Abfolge oder Sequenz der Aminosäuren bestimmt gemeinhin die dreidimensionale Struktur des Proteins, welche wiederum eng mit der Proteinfunktion verflochten ist.

Durch das Studium verschiedener Aspekte von Proteinen konnten bereits wertvolle Erkenntnisse über ihre Dynamik, Interaktionsmechanismen und Strukturen gewonnen werden. Was jedoch weitgehend unverstanden bleibt, ist der Zusammenhang von Aminosäurensequenz und Proteinstruktur. Wie findet eine unstrukturierte Polypeptidkette innerhalb kürzester Zeit die eine Konformation, in welcher sich die stabile dreidimensionale Struktur des Proteins manifestiert?

Die Vorhersage einer Proteinstruktur anhand ihrer Aminosäurensequenz ist mitunter relativ erfolgreich, sofern Strukturen homologer, d.h. sequenzverwandter Proteine, bereits bekannt sind. Diese, auch strukturelle, Verwandtschaft resultiert aus der Evolution der Proteine, in der möglicherweise wenige ursprüngliche Strukturelemente in mannigfaltigen Anordnungen wiederverwendet wurden.

Die Natur scheint ein vergleichsweise kleines Repertoire an Proteinarchitekturen zu benötigen um alle erdenklichen Funktionen abzudecken. Zusammen mit der evolutionsbedingt oft hohen Sequenzähnlichkeit innerhalb einer Proteinarchitektur, gibt dies Grund zur Annahme, dass ein Grossteil des Sequenzraums, d.h. der theoretisch möglichen Aminosäuresequenzen, nie von der Natur verwendet wurde und in diesem Sequenzraum eventuell noch viele weitere stabil faltende Proteine vorkommen.

Eine Hauptmotivation dieser Doktorarbeit war es, durch die Suche nach grundlegend neuartigen Proteinen, die stabil falten, zur Entschlüsselung des Proteinfaltungsproblems beizutragen und dabei unter Umständen das Repertoire an Proteinarchitekturen um noch nicht da gewesene Bauarten zu erweitern.

Für die Suche nach grundlegend neuartigen Proteinen lassen sich zwei entscheidende Faktoren feststellen. Zum einen bedarf es einer hochwertigen Bibliothek, welche sich im nicht erschlossenen Sequenzraum erstreckt. Zum anderen benötigt man ein Selektionssystem, welches möglichst direkt auf kompakte Proteinfaltung prüft. In dieser Arbeit wird beiden Faktoren einzeln und zusammen im Verbund nachgegangen.

Von den beiden Bibliotheken, die in den Selektionen auf kompakte Proteinfaltung benutzt wurden, war die erste eine bereits etablierte Bibliothek, bestehend aus kleinen, kombinatorisch angeordneten, "binary patterned" Bausteinen für Sekundärstrukturelemente, ohne signifikante Ähnlichkeit zu natürlichen Proteinen. Diese Bibliothek aus Sekundärstrukturbausteinen war gekennzeichnet durch eine sehr heterogene Grössenverteilung, welche hauptsächlich auf die hohe Sequenzähnlichkeit innerhalb der vorkommenden Varianten der jeweiligen Bausteine zurückzuführen war. Durch Optimierung der Aufreinigung und Amplifikation konnte die heterogene Grössenverteilung beträchtlich korrigiert werden. Schliesslich konnte durch die

Kombination dreier Bausteine, der Bibliothek aus Sekundärstrukturbausteinen, einem neuen Baustein für fünf aufeinander folgende hydrophobe Aminosäuren sowie, ein weiteres Mal, der Bibliothek aus Sekundärstrukturbausteinen, eine neue Bibliothek geschaffen werden, die für rund 100 Aminosäuren kodiert.

Die zweite Bibliothek, die in der Suche nach grundlegend neuartigen Proteinen durch Selektionen auf kompakte Proteinfaltung benutzt werden sollte, wurde von Grund auf neu geschaffen. Dabei handelt es sich um eine komplett zufällige Bibliothek, ohne Einschränkung der Art oder Position der Aminosäuren, bestehend aus NNK Codons sowie einem zusätzlichen zentralen Modul, welches wieder fünf aufeinander folgende hydrophobe Aminosäuren umfasst. Dies ist nach unserer Kenntnis die erste zufällige Bibliothek mit nahtlos verbundenen Modulen. Der nahtlose Aufbau wurde erst möglich durch den Einsatz von Restriktionsenzymen des Typs IIS, welche DNA-Überhänge mit zufälliger Sequenz erzeugten, sowie durch Anfügen konstanter DNA-Blöcke unterschiedlicher Länge an die Module mit zufälliger Sequenz. Zu guter Letzt konnten durch spezifische Erkennungssequenzen ausschliesslich die gewünschten Modulkombinationen aus den diversen Ligationsprodukten amplifiziert werden.

Ein Selektionssystem, welches möglichst direkt auf kompakte Proteinfaltung prüft, könnte in Form des bakteriellen Tat-Membrantransports gefunden werden. Dieser vollbringt den Export komplett gefalteter Proteine durch die Zellmembran und besitzt eine Qualitätskontrolle für korrekte Faltung, wodurch ungefaltete Tat-Substratproteine vom Export ausgeschlossen werden.

Anhand einer Palette von Proteinen unterschiedlicher Faltungsfähigkeit und mithilfe des Enterobakteriums *E. coli* wurde die Effizienz des Tat-Membrantransports ermittelt und mit den beiden durchsatzstärksten Proteintransportwegen ins Periplasma, via SecB bzw. SRP, verglichen.

Um eine quantifizierbare Betrachtung zu ermöglichen, inwieweit der Tat-Membrantransport eine kompakte Faltung bedingt, wurden mit der β-Laktamase und GFP zwei verschiedenartige Reporterproteine herangezogen.

Über die β-Laktamase war bereits bekannt, dass sie als Reporterprotein eine Korrelation von Proteinlöslichkeit und Tat-Membrantransport herstellen kann. Die β-Laktamase wurde hier hinsichtlich ihrer Eignung untersucht, ob und in welchem Mass sie als Tat-Reporter infrage kommt für die Unterscheidung von löslichen, aber unterschiedlich gut gefalteten Proteinen.

Darüber hinaus wurde mit GFP ein leistungsfähigeres Tat-Reporterprotein etabliert, welches sich dadurch auszeichnet, dass es keinen zusätzlichen Überlebensdruck auf die Bakterien ausübt und keine oder nur geringe periplasmatische Fluoreszenz besitzt, wenn es über einen anderen Weg als den Tat-Membrantransport exportiert wird. Allerdings muss bei GFP sichergestellt werden, dass jegliche Fluoreszenz, die sich nicht im Periplasma befindet, effizient beseitigt wird. Dabei sollte die Tat-Membrantransportrate mit ihrer vorgeschalteten Faltungskontrolle nicht beeinträchtigt werden, um ein möglichst starkes, exklusiv periplasmatisches Signal zu erhalten, welches seinerseits mit den Faltungseigenschaften des an GFP fusionierten Proteins in Zusammenhang stehen sollte.

Um die geeignete Balance zwischen möglichst hoher Tat-Membrantransportrate und effizienter Beseitigung nicht periplasmatischer Proteine, fusioniert an den Reporter GFP-ssrA, zu finden,

wurden verschiedene Bestandteile des cytoplasmatischen ClpXP Hydrolasesystems moduliert und abgeschwächte Varianten des C-terminalen ssrA Markierungspeptids getestet, welches die Proteine für den Abbau durch ClpXP kennzeichnet.

Mithilfe einer Palette von Testproteinen wurden die Parameter beider Reportersysteme letztlich soweit verfeinert, dass sie zusammen mit den beiden Bibliotheken als Tat-Selektionssystem für grundlegend neuartige, stabil gefaltete Proteine eingesetzt werden konnten.

Eine Vielzahl an Selektionsrunden brachte bislang noch keine stabil gefalteten Proteine hervor. Gleichwohl zeigte der Einfallsreichtum der entdeckten, unvorhergesehenen und aussergewöhnlichen, bakteriellen Umgehungsmechanismen gewisse Grenzen der benutzen Tat-Selektionssysteme auf, ermöglichte dadurch aber auch eine weitere Optimierung ihrer Leistungsfähigkeit. Einige der beobachteten Umgehungsmechanismen sprechen stark dafür, dass die benutzten Tat-Selektionssysteme voraussichtlich durch andersartige Selektionssysteme, die nicht auf einem Membrantransport basieren, ergänzt werden müssen, um gewisse Ausprägungen falsch-positiver Signale auszuschliessen.

Obgleich die hier vorliegenden Ergebnisse die Komplexität des Proteinfaltungsproblems abermals deutlichmachen, hoffen wir, mit der Herstellung der Bibliotheken, die sich im nicht erschlossenen Sequenzraum erstrecken, sowie mit der Etablierung der leistungsstarken Tat-Selektionssysteme, die nahezu direkt auf kompakte Proteinfaltung prüfen, entscheidend zu der Suche nach grundlegend neuartigen Proteinen beitragen zu können.

Abstract:

Proteins are essential for most aspects of living cells. These fascinating macromolecules are fundamental constituents of life as we know it. The amino acid sequence of a protein normally determines its three dimensional structure, which is often closely coupled to the function of the protein.

Great advances have been made in the comprehension of proteins, regarding their dynamics, interaction mechanisms, and structures. Yet, the principles of protein folding by which an unstructured polypeptide is able to quickly adopt its stably folded conformation are still not fully understood.

Prediction of the three dimensional structure of a protein from its amino acid sequence is sometimes possible with a certain precision if structures of proteins with homologous sequences are available. This is due to the evolution of proteins, possibly originating from few structural elements, which have been re-used in diverse arrangements.

Natural proteins are able to fulfill their various functions using a comparatively small number of protein folds, which suggests that an extensive part of protein sequence space has not been sampled by nature and may contain many more stably folding sequences.

This work was driven by the desire to contribute to a better understanding of the principles governing protein folding by seeking truly novel, stably folded proteins and thereby extending the perspective beyond protein folds found in nature.

When looking for truly novel proteins by experimental screening, two factors are crucial: A valuable library that covers unexplored sequence space and a selection system that targets protein folding as directly as possible. In this thesis both issues are addressed individually and in combination.

One of the two libraries used for selections towards well-folded proteins was a previously established combinatorial library, composed of binary patterned modules for secondary structure elements without significant homology to natural proteins. This secondary structure library showed a very heterogeneous size distribution, which could largely be attributed to the high intra-module sequence similarity. The size distribution was partially resolved by improving purification and amplification procedures. A library encoding about 100 amino acids was obtained by incorporating a module that encodes a patch of five consecutive hydrophobic amino acids in-between two entities of the secondary structure library.

The second library used for selections towards truly novel, well-folded proteins was created from scratch as a completely unbiased, fully random library composed of NNK codons with the addition of a central module encoding a hydrophobic patch. To our knowledge, it is the first random library where the modules are seamlessly assembled. This was accomplished by using type IIS restriction enzymes that generate overhangs in the randomized sequence, fusing the construction modules to DNA-stretches of distinctive lengths, and incorporating signature sequences that allow the exclusive amplification of the desired ligation product.

A selection system that is able to address protein folding directly might be achievable using the bacterial Tat pathway, which accomplishes the translocation of completely folded proteins and contains a folding quality mechanism that rejects improperly folded Tat substrates.

Using *E. coli* and a set of test proteins, the translocation efficacy of the Tat pathway was compared to the SecB and SRP pathways, two major export pathways to the periplasm.

To quantify the correlation between Tat-dependent translocation and the folding properties of the proteins targeted to the Tat pore two distinct reporter proteins were used. The potential of β-lactamase as reporter was studied, regarding its discrimination of folding states beyond a previously described Tat-dependent selection based on protein solubility. Additionally, a more Tat-exclusive reporter system without survival pressure was established by using GFP as reporter.

To correlate GFP fluorescence with Tat-dependent export, the non-periplasmic GFP signal had to be eliminated at a rate that still allows sufficient translocation via the Tat pathway, including the folding quality control. The different components of the ClpXP-dependent cytoplasmic degradation of proteins fused to GFP carrying an ssrA or similar tag were modulated in several aspects in order to balance Tat-dependent translocation and removal of non-periplasmic fluorescence.

Using a set of test proteins, the parameters for both reporter setups could be refined to attain Tat-dependent selection systems closely coupled to protein folding, which were then used in combination with the two libraries for selections towards truly novel proteins.

The outcome of the numerous selections did not yield stably folded proteins so far. Still, the wealth of unexpected and astonishing escape mechanisms observed did provide valuable insights into the limitations of these selection setups and helped to further optimize their performance. Some of the observed escape mechanisms made evident that these Tat-dependent selection setups probably need to be complemented by methods not based on translocation, in order to eliminate certain types of false positives.

These results may emphasize the complexity of the folding problem. However, with the construction of libraries that cover unexplored sequence space and the development of a potent selection system that targets stable protein folding we hope to provide promising tools to be used in the search for truly novel, well folded proteins.

1. Introduction

1.1 Protein fold space and sequence space

Proteins are fascinating macromolecules. Without proteins, life as we know it would not be possible, as they are essential for almost all biological activities of a cell, especially metabolism. The repertoire of their functions includes energy conversion and enzymatic activity catalyzing many chemical reactions, transport and relay processes, as well as structural support and specific recognition and binding of molecular shapes.

A protein's function is generally dependent on its structure. The protein structure is determined by its amino acid sequence [1] and normally represents the global free-energy minimum of this sequence in the given environment.

The mechanisms and forces of folding an unstructured polypeptide with gigantic degrees of freedom into a unique structure in a time frame of milliseconds to seconds have captivated researches for a long time [2,3] and are still not fully understood. Protein folding is a transition from disorder to order. A current working model proposes that proteins have funnel-shaped energy landscapes with many conformations at high-energy states and increasingly fewer at low-energy states as folding proceeds [4-6]. This changing density of states is linked to conformational chain entropy or more generally to intramolecular-plus-solvation free energy. A rugged funnel can describe conformational heterogeneity, multiple folding pathways, influences of folding conditions on the dominant folding route and provide a microscopic framework for folding kinetics.

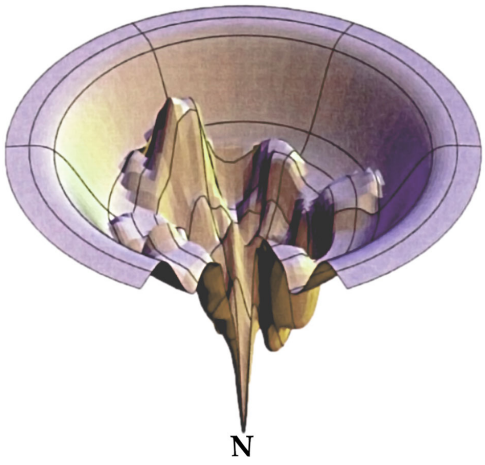


Figure 1.1: A model of a funnel-shaped rugged energy landscape with kinetic traps, energy barriers, and some narrow throughways to the native state (N), adapted from [5].

One hypothesis of a general funnel-shaped folding mechanism is zipping and assembly [7]. In zipping and assembly short fragments of the polypeptide chain independently search for local metastable structures. Metastable local structures zip to form larger, more stable structures, which then assemble further and further till metastability becomes stability and the end of the funnel is reached, thereby greatly reducing the searched conformational space. This mechanism is

especially useful in computational approaches, where a global optimization process, seeking the native state of a protein using merely physical models, poses a needle-in-a-haystack problem.

Structural biology has by now determined a huge number of the structures of natural proteins and the data has been made freely and publicly available in the Protein Data Bank (PDB) [8,9].

Even early on, with comparatively few structures available, classification of protein folds was already pursued [10]. Today, SCOP [11] (structural classification of proteins) and CATH [12] (class, architecture, topology, homology) represent the two most renowned catalogs of protein folds. A protein fold can generally be described as ordered three-dimensional arrangement of secondary structure elements constituting a protein domain. Classification of protein folds is complicated since the boundaries of folds may seem arbitrarily and inconsistently defined and protein-fold classifications differ from one another [13]. Nevertheless, the general scheme for clustering protein structures is consented [14,15].

Even though the number of protein folds found in nature depends on the parameters of classification [16] and might range from 1'000 [17] to 10'000 [18-20]. It appears that nature uses only a comparatively small number of protein folds.

Comparing sequence pairs forming the same fold often shows high similarity [21-23], but several folds contain sequence pairs with little homology [24], e.g. in the SCOP superfamily annotation.

The other extreme also exists as metamorphic proteins in nature [25-27] and more radical in engineered proteins [28], where similar sequences adopt different folds. The difference in energy between the two folds for a given sequence is low and a switch can be induced by few amino acid mutations [29,30] or changes in the local environment. Prions represent an irreversible switch with a continuous formation of fibrils, as many switched proteins assemble to few fibrils.

Regarding the evolution of today's stable proteins, different models have been proposed about the onset of common structural prototypes (single birth [31] vs. multiple birth model [32]). Irrespective of the course of events, the solved structures show an interesting (power-law) distribution, where a large percentage of the protein families assume one of few popular folds (superfolds) and an estimated 80% of all protein families adopt one of about 400 mesofolds, and the remaining majority of unifolds being covered just by single families [20,33]. Concordant with their large number of variants, the superfolds show high tolerance towards mutations [34,35]. These superfolds may have been established early in the evolution of proteins and thus have had the most time to radiate in sequence space, or alternatively they may represent physically optimal folds [36] and have the highest propensity to arise through random evolution [18].

The evolution of proteins and possibly also their folds is characterized by the re-use of once established elements by duplication and mutation, (circular) permutations and insertion or deletion [13,33].

As sequence space is astronomically large compared to the known fold space, the propensity of discovering a spot where the sequence adopts e.g. a stable supersecondary structure may be so low that established elements are re-used as often as possible. Looking at the superfolds and

mesofolds, it seems nature can fulfill most biological tasks by using a small repertoire of folds. On the other hand, the large number of unifolds, which may be of more recent origin, also suggests that there might be many additional spots in sequence space, which facilitate a stably folded structure.

Computational approaches [37,38] suggest as well that far more stable protein folds might exist. This raises the fundamental question underlying this thesis: how can we find truly novel proteins?

1.2 How to find truly novel proteins

The information from solved protein structures can be employed for predicting the structure and function of homologous sequences [39], as protein folds are related and structural elements are re-used in various arrangements. Conversely the amino acid sequence for a designed structure of a novel fold can be deduced from known structures, e.g. by assembling short peptide fragments adopting the desired secondary structures in the PDB [40,41]. Simultaneous optimization of sequence and structure using the Rosetta program [42], allowing small deviations from the initial design, resulted in an extremely stable protein of 93 amino acids, named Top7. The X-ray crystal structure of Top7 (PDB accession code 1QYS) showed a striking similarity to the design model at atomic resolution (1.17 Å RMSD over all backbone atoms) [43]. However, it has to be emphasized that this proof of concept example Top7 has so far been the only one reported.

Another successful approach for finding novel proteins came from a completely different direction. Novel ATP-binding proteins of 80 amino acids could be discovered by functional selections on a large random sequence library [44]. Selections for ATP binding were performed in vitro by mRNA display for 18 rounds, with error-prone PCRs increasing the diversity in round 10-12. The protein-ligand-complex crystal structure of this first protein derived from a random library by in vitro evolution revealed a novel fold and an unprecedented stabilization by a zinc ion [45], which had not been a consideration in the original library.

Generalizing the second approach, the experimental selection of novel folded proteins from a diverse library, identifies two key criteria in the search for truly novel proteins by non-computational means: a library that encodes proteins with very low homology to natural proteins and a suitable selection system that is likely to yield folded proteins under physiological conditions.

1.2.1 Library designs for encoding novel proteins

In most cases, protein domains consist of 50 to 200 residues [46-48]. Yet, a fully random polypeptide with a length of already eleven amino acids has a theoretical diversity of 20^{11} or $\sim 2\times10^{14}$, which is beyond the limit that can be reasonably handled in wet-lab experiments [49].

Depending on the length of the encoded protein, the theoretical diversity of a library can thus quickly exceed the diversity that can be screened experimentally by many orders of magnitude. To circumvent this sampling problem, protein libraries with predefined properties and a reduced theoretical diversity can be envisioned [50].

A secondary structure library (SSL), which was constructed in our laboratory, is based on the binary patterning principle of polar and non-polar residues [51,52] and was built without restricting the topology to a known fold. This combinatorial library encodes for proteins of an average length of 100 amino acids, composed of random arrangements of modules for secondary structure elements (α-helix, β-strand, and β-turn) [53]. Proteins encoded by the SSL show no significant homology to natural proteins. Arbitrary members composed of mainly α-helical modules showed signs of α-helical secondary structure, a defined oligomerization state, but rather a molten globule behavior than a defined structure. Proteins composed of mainly β-modules formed highly ordered amyloid-like aggregates.

A revised version of this library, SSL2.1, was constructed, where the pI-distribution is more similar to naturally occurring proteins and independent of the turn-module content. Furthermore, the SSL2.1 could be constructed at a experimental diversity of 10^{12}, 1'000 times higher than the prototype SSL [54]. The SSL2.1 was utilized for this thesis.

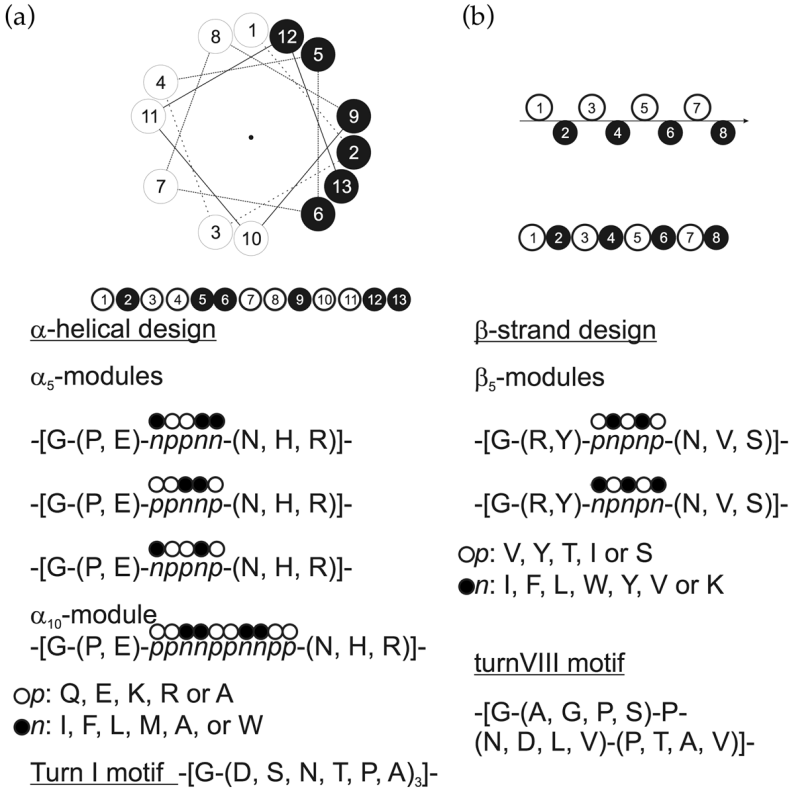


Figure 1.2: Modules of the SSL2.1 with their binary patterning, adapted from [54].
(a) α-helical design and turn I motif, (b) β-strand design and turn VIII motif. Empty circles and "p" represent polar residues, filled circles and "n" represent non-polar residues. Amino acids in one letter representation, square brackets [] delimit module boundaries, parentheses () describe a set of residues for one position, e.g. $(D,S,N,T,P,A)_3$ would be represented as [DSNTPA]{3} using regular expressions.

Although the idea to bias a library in favor of folded (secondary) structures sounds compelling, binary patterning has been successful mainly in the design of four-helix bundles [51,55]. β-sheet proteins based on binary patterning are often aggregation prone as they form huge intermolecular oligomeric assemblies that look like amyloid fibrils [53,56,57]. The chosen bias in library design

may thus hinder the discovery of other stable folds, since their putative sequence space is not covered by the library.

The use of completely random sequences does not restrict the covered sequence space in any way but also poses the biggest sampling problem, as the theoretical sequence space becomes gigantic. See 2.3 for further introduction.

1.3 Selection systems towards folded proteins

As direct selection for folding is not (yet) possible, other characteristics of well-folded protein are exploited in selections towards folded proteins. Well-folded natural proteins mostly adopt an intramolecular compact conformation, possess excellent solubility in the cytosol, and their function often depends on the defined tertiary structure. Function, solubility and compactness are the main properties that can be used as selection criteria towards well-folded proteins.

Functional selection is often carried out as selection for binding molecular shapes with high specificity or affinity. Depending on the selection system, this is often correspondingly coupled to selection for solubility, as aggregating proteins are not present in the output of the selection. The ATP-binding protein from a random library was obtained by functional selection for ATP binding [44] by mRNA display [58]. The washing steps in in vitro selection schemes remove non-binding or aggregating proteins and their genotypes from the selection pool.

Selection for solubility can be carried out in many different ways: in vivo as fusion or insertion within a variety of reporter proteins [59] or in vitro, where it is mostly driven by methods that target protein compactness and insoluble proteins are normally removed as intrinsic aspect of the experimental procedure.

Compact folding of a polypeptide into one defined tertiary protein structure is difficult to target directly. Therefore, selections for compactness are approached tangentially from different directions. One of the most successful methods is selection for protease resistance coupled with phage display [60-63] or ribosome display [64]. Other attempts to select for compactness include insertion of libraries into a loop of a binding domain coupled with selections for recovered binding [65,66] and the removal of aggregation-prone proteins with exposed hydrophobic regions by hydrophobic chromatography [64].

In vitro methods generally possess the advantage that a considerably higher diversity of different proteins can be sampled, up to 10^{12} variants, compared to about 10^{9} for in vivo selections.

However, selections on the SSL using phage display or ribosome display resulted mainly in short peptides or other escape mechanisms, e.g. avoiding protease-targets in the primary structure, that do not show properties of well folded proteins [54].

A selection system, which is able to address protein folding directly, might rather be found in vivo, where sophisticated molecular mechanisms have evolved, e.g. in the secretory pathway of yeast [67] or bacteria.

1.4 Protein translocation systems of *E. coli*

As Gram-negative bacterium, *Escherichia coli* contains an inner membrane that separates the cytosol from the periplasmic space, which in turn is confined by an outer membrane. The inner membrane hosts three major [68] protein transport systems that facilitate membrane insertion and periplasmic translocation of proteins carrying appropriate signal sequences: the Sec translocon, the YidC insertase, and the Tat system [69] (**Figure 1.3**).

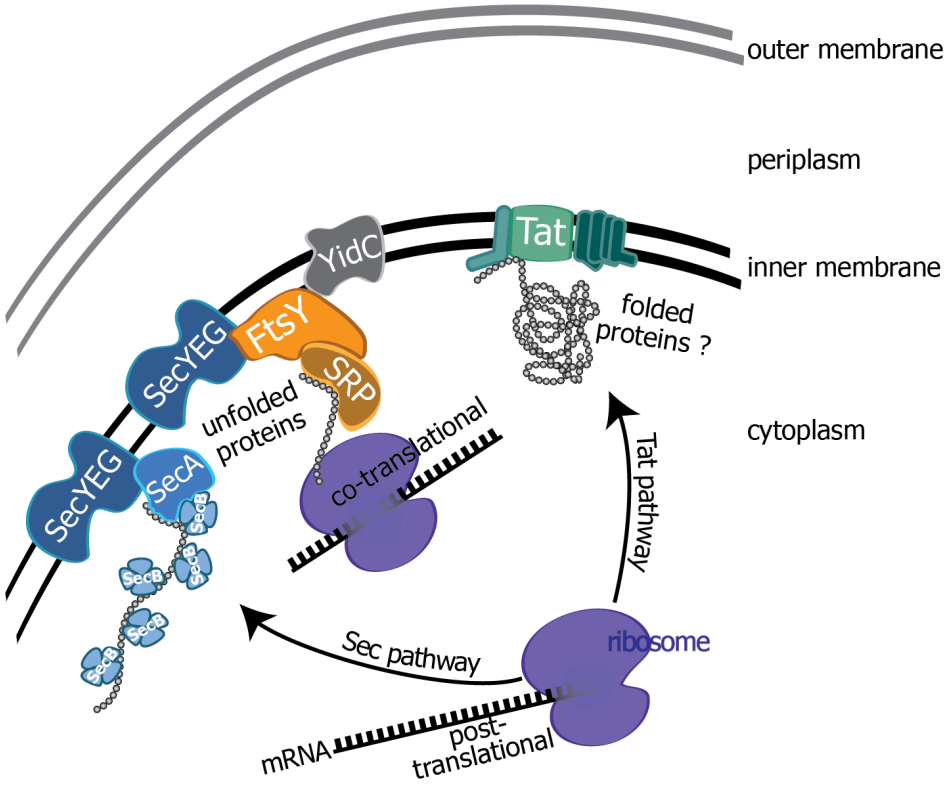


Figure 1.3: Schematic depiction of the three major protein translocation pathways in *E. coli*.

The heterotrimeric Sec translocon is the best characterized translocation system and essential for viability. It is composed of the subunits SecY, SecE and SecG. The membrane-embedded protein-conducting channel SecYEG is probably the major protein transport site for the transport of secretory proteins across the inner membrane and insertion of membrane proteins into the inner membrane. The Sec pathway can be used post-translationally, where the pre-proteins are fully translated in the cytoplasm or co-translationally, where transport occurs concurrently to protein translation on the ribosome.

In the mainly post-translational branch of the Sec pathway, the nascent pre-proteins are recognized directly by the cellular chaperone SecB [70]. It maintains secretory pre-proteins in a translocation-competent state and interacts specifically with the membrane-bound motor protein SecA, which is associated with SecYEG. The tetrameric SecB chaperone is important for the export of several secretory proteins but is not essential for viability [71].

In the other, mainly co-translational branch of the Sec pathway, the ribonucleoprotein signal-recognition particle (SRP) binds the ribosome-nascent chain complex. Translation in the

cytoplasm is halted and the whole complex is directed to the membrane-anchored SRP-receptor FtsY, which is associated with SecYEG. On SecY, FtsY uses the same binding site as SecA. The SRP pathway is primarily used for the insertion of proteins spanning the inner membrane. Yet, secretory proteins may be routed via the SRP pathway, if they carry long and very hydrophobic signal peptides.

The N-terminal SecYEG-targeted signal sequences of secretory proteins begin with a positively charged N-region, followed by a central hydrophobic H-region and a polar C-region that harbors the signal peptidase cleavage site, where the signal sequence is cleaved off after transport. Signal sequences for membrane proteins mostly lack signal peptidase cleavage sites as they contain more hydrophobic stretches that can become integrated in the membrane or serve as anchor.

Alternatively, the YidC insertase, a multi-spanning membrane protein, can integrate membrane proteins into the inner membrane independently. It can also associate with SecYEG to facilitate the release of transmembrane regions from the SecY channel. SRP can bind to the YidC insertase directly and the targeting of many YidC substrates is dependent on the SRP pathway.

In contrast to the other bacterial transport systems, where targeted proteins are kept in an unfolded state or transport occurs co-translationally, the Tat system accomplishes the transmembrane passage of completely folded proteins or protein complexes. It consists of three proteins in the inner membrane, TatA, TatB and TatC. The mechanisms by which fully folded Tat substrates that often contain co-factors can pass the membrane are still poorly understood [72].

1.5 Properties of the Tat pathway, a potential in vivo selection system for folding

The Tat pathway was named after the twin Arginine residues in the consensus motif S-R-R-x-F-L-K (x being a polar amino acid) of Tat signal sequences [73,74]. The consensus motif is located close to the end of their N-region, followed by a central hydrophobic H-region and a polar C-terminal region with a signal cleavage site. Tat signal sequences thus have a similar tripartite composition as Sec signal sequences and are also cleaved by signal peptidase I [75,76]. Yet, Tat signal sequences are often longer than Sec signal sequences, have a lower net hydrophobicity [77,78], and carry positively charged residues proximal and distal to the signal peptide cleavage site, which function as a Sec avoidance signal [79,80].

The twin Arginine pair of the signal-sequence consensus motif is almost invariant and only few substitutions of one Arginine to Lysine, Asparagine or Glutamine are tolerated without abolishing Tat-dependent translocation [81-83].

Redox proteins of the anaerobic respiration machinery and proteins for the biogenesis and remodeling of the cell envelope make up a big portion of the natural bacterial Tat substrates. Most proteins transported via the Tat pathway are secreted, although Tat substrates that contain a membrane anchor have been reported as well [84].

Proteins targeted to the Tat translocase generally fold stably in the cytosol, acquire and incorporate co-factors needed for maturation, or form quaternary structures by oligomerization

[85-88]. Co-translocation of protein complexes via the Tat pathway can also occur by a hitchhiker mechanism, where only one subunit contains a Tat signal sequence [89-91].

The typical Tat translocase of *E. coli* assembles from TatA, TatB and TatC. Additionally, the TatA paralogue TatE can be expressed and substitute for TatA [92].

TatC is a polytopic membrane protein with six transmembrane domains, whereas TatA and TatB are single spanning membrane proteins [93,94].

TatA is an 89 amino acid protein consisting of an N-terminal transmembrane helix, a short hinge region, an amphipathic helix, and an unstructured C-terminal region. TatB consists of 171 amino acids, shows 20% sequence identity to TatA, and shares the same modular structure plus a longer C-terminal part. TatC consists of 258 amino acids and folds into six transmembrane helices. The first crystal structure of a TatC paralogue from Aquifex aeolicus depicts that "the transmembrane helices form a curved wall overhung on the concave face by the periplasmic cap" [95]. The periplasmic cap, formed by the first two periplasmic loop regions as well as the cytosolic N-terminus, and the first cytosolic loop are deemed essential for activity.

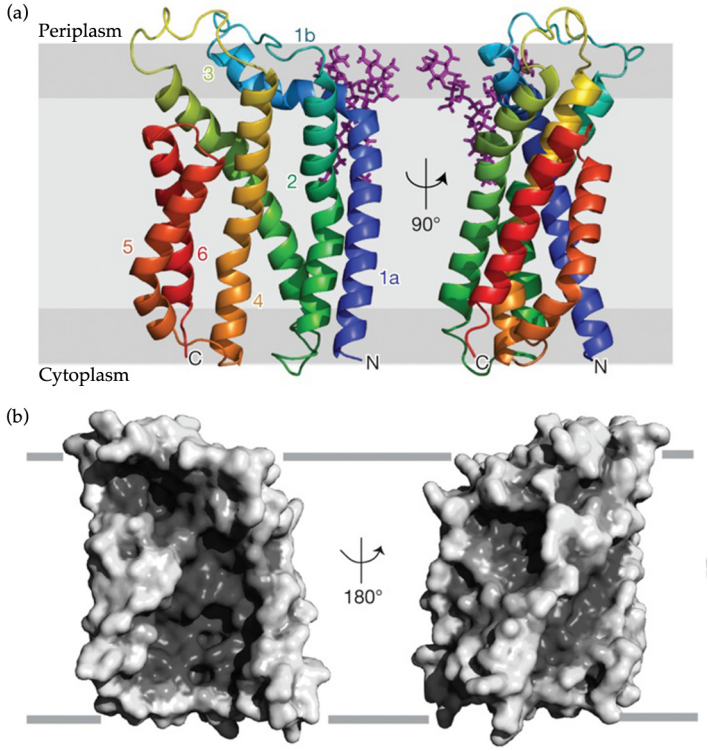


Figure 1.4: Structure of TatC from the hyperthermophilic bacterium Aquifex aeolicus [95]. (a) Cartoon representation colored from blue at the N terminus to red at the C terminus. A semi-ordered molecule of LMNG detergent is present and is shown in purple. (b) Surface representation. The left hand views in (a) and (b) are the same orientation.

The current working model [69] of the processes in the translocation via the Tat pathway can be summarized in the following simplified scheme. The consensus motif of the Tat signal sequence is first recognized by TatC, then the whole signal peptide is bound in a binding pocket formed by a TatBC complex, which can consist of multiple TatBC heterodimers. TatC possesses a signal peptide insertase activity and probably drives the loop-like insertion of the signal sequence

between transmembrane helices of TatC and TatB. The remaining part of the translocation is solely dependent on the proton-motive force as energy source. Numerous TatA protomers are recruited to the complex and constitute the functional Tat translocase by a pore forming or membrane-destabilizing effect.

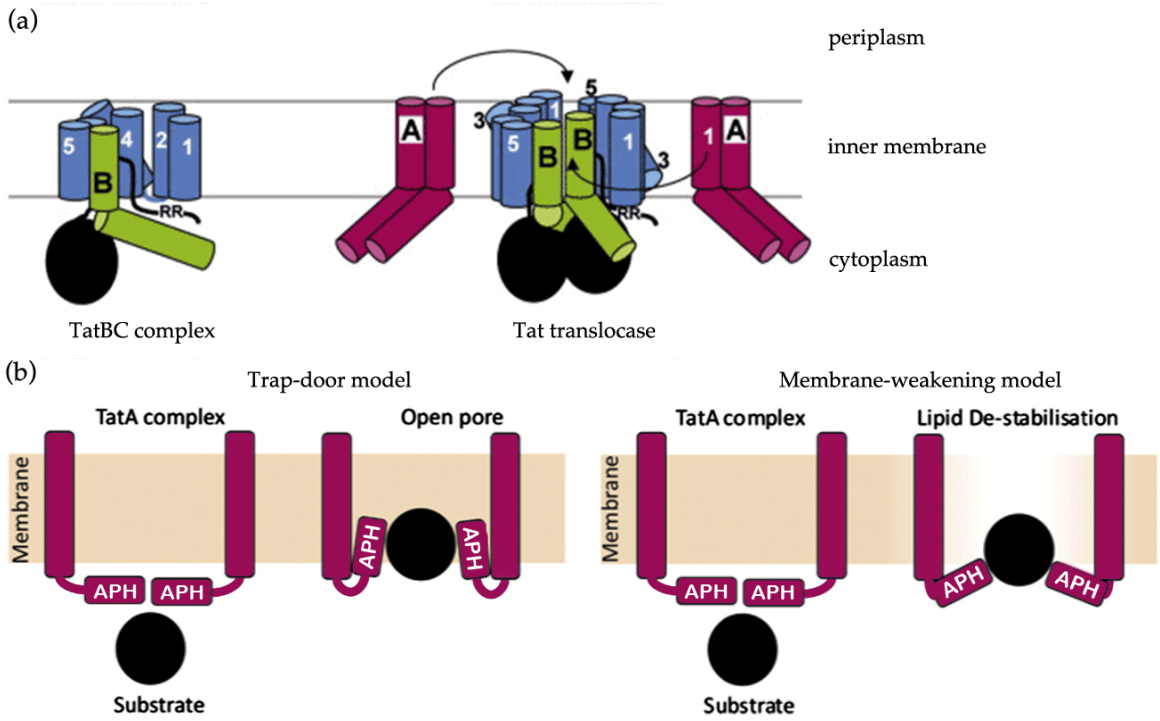


Figure 1.5: Model of the Tat-dependent translocation mechanism.
(a) Binding of a Tat substrate to the TatBC receptor complex, several precursor-TatBC-complexes assemble to an oligomeric complex that recruits TatA protomers [69],
(b) Two proposed models for the subsequent translocation event involving the TatA complex. Trap-door model and Membrane-weakening model [96].

However, the feature that could render the Tat translocase an effective tool towards selection of well folded proteins, namely the molecular nature of the folding quality check rejecting improperly folded Tat substrates, is not well understood. Tat-dependent proteins that undergo co-factor insertion could require redox enzyme maturation proteins as proof-reading chaperones before targeting the Tat translocase [97-100]. Yet, no cytosolic chaperones seem to be required for the export of cofactor-less Tat substrates [101]. Mutational studies on TatBC strongly suggest its direct participation in the structural proofreading of substrate proteins [102]. Unfolded proteins with a Tat signal sequence may also bind to the Tat translocase [86,103]. The proofreading mechanism is presumed to reject the translocation and route the unfolded protein back to the cytosol or even to initiate the degradation of rejected molecules at the membrane or in the cytosol [104,105].

Notably, unfolded polypeptides of up to 120 amino acids might be able to bypass the proofreading and get exported via the Tat translocase. Tat export of unstructured polypeptides is efficiently blocked when an exposed hydrophobic patch is present [106]. This has to be taken into specific consideration when designing a Tat selection system for folded proteins of about 100

amino acids, although Tat-dependent translocation may also be influenced by the fusion of a stably folded reporter protein.

Thus, the Tat pathway may have great potential as a selection system for folding if certain prerequisites are fulfilled. The presence of a stretch of hydrophobic amino acids in the proteins of interest could preclude the translocation of unfolded states, and a reporter protein linked to the proteins of interest could facilitate to quantitatively correlate the Tat-dependent translocation with a phenotypic readout.

Well established reporter proteins for solubility [59] can essentially be used, provided that their presence in the periplasm can be detected unambiguously from cytoplasmic expression.

1.5.1 β-lactamase as reporter, Tat-dependent selection for solubility

An assay that correlates Tat-dependent translocation with solubility of the proteins of interest was previously reported [107]. In this assay a C-terminally fused β-lactamase serves as reporter protein. After translocation to the periplasmic space, it confers resistance to β-lactam antibiotics, which is used as readout. This assay could correlate the solubility of the proteins of interest with cell survival under selective conditions by using the folding quality control of the Tat pathway. It was successfully used as selection system for solubility-enhanced variants of the Alzheimer's Aβ42 peptide.

Moreover, β-lactamase was used as reporter protein in assays of Tat-dependent hitchhiker translocation of strongly interacting proteins [90,91].

It has to be noted though that functional β-lactamase can be released from the cytoplasm by other mechanisms, which may interfere with Tat-dependent selection for solubility or folding.

β-lactamase is translocated at significantly higher rate via the Sec pathway, if a peptide stretch is present, which is recognized as Sec signal sequence. Most organisms employ a Sec-dependent translocation of β-lactamase, and the Tat pathway in general has a much lower transport efficacy [108,109].

Even without detectable signal sequence, bare β-lactamase was reported to translocate to the periplasm at low rates [110]. This could occur when the Tat signal sequence and an unfolded polypeptide are proteolytically cleaved of the folded β-lactamase in the cytoplasm.

Expression of certain toxic constructs may further lead to cell lysis and release of functional β-lactamase from the cytoplasm.

Therefore, Tat-dependent selections towards well folded proteins may need to be complemented by other reporter proteins whose functional presence in the periplasm is exclusively dependent on translocation via the Tat pathway.

1.5.2 Export of fluorescent GFP and cytoplasmic degradation by ClpXP

The green fluorescent protein (GFP) [111] of the Pacific North-west jellyfish Aequorea victoria has been widely used as reporter protein. In *E. coli*, it was shown to be fluorescent in the periplasm only when translocated via the Tat pathway [109,112-114], after folding and

establishment of the fluorophore by autocatalytic intramolecular cyclization in the cytosol. Yet, this implies that GFP fluorescence is present in the cytoplasm and a fluorescent readout of the whole cell cannot be easily correlated with Tat-dependent translocation. And although redox-sensitive variants of GFP have been engineered [115,116], which would allow a discrimination of localization in the oxidizing cytoplasm versus reducing periplasm, they differ mainly in their excitation maxima and have similar emission profiles. The flow cytometry technologies available today cannot be used to detect exclusively the periplasmic fraction of redox sensitive GFP as both fractions have the same emission maxima.

Selective and efficient degradation of cytoplasmic GFP by proteolysis would result in a GFP fluorescence signal that is mainly dependent on periplasmic localization via the Tat pathway. Thereby whole-cell fluorescence could be correlated with Tat-dependent translocation. And provided that bypassing the Tat proofreading mechanism can effectively be impeded, the translocation will be directly linked to the stable folding of the substrate. Thus, the total fluorescence signal would be quantitatively coupled to the folding behavior of the Tat-targeted proteins.

Furthermore, a robustly folded version of GFP (superfolder GFP) was engineered that folds well even when fused to poorly folded polypeptides, and exhibits higher fluorescence signals. It also shows improved folding kinetics and increased resistance to denaturation [117].

A system that targets cytoplasmic proteins for specific and rapid degradation in *E. coli* and many other organisms is the ssrA/ClpXP machinery. It was reported to efficiently eliminate the cytoplasmic fraction of a Tat-targeted GFP-ssrA fusion and only secretion via the Tat pathway rescued the protein from cytoplasmic proteolysis by ClpXP, conferring a fluorescence signal proportional to the amount of translocated GFP [118].

The stacked ClpXP protease complex consists of two distinct proteins; ClpX forms a hexameric AAA+ (ATPases associated with diverse cellular activities) unfoldase, whereas ClpP is active as a 14-subunit self-compartmentalized peptidase. The axial pore of the asymmetric hexameric ClpX ring binds unstructured regions of the substrate proteins. Powered by continuous ATP hydrolysis the substrate is unfolded and pulled trough the ClpX pore and the unfolded polypeptide it translocated into the closely attached proteolytic chamber of ClpP, where it is degraded [119,120].

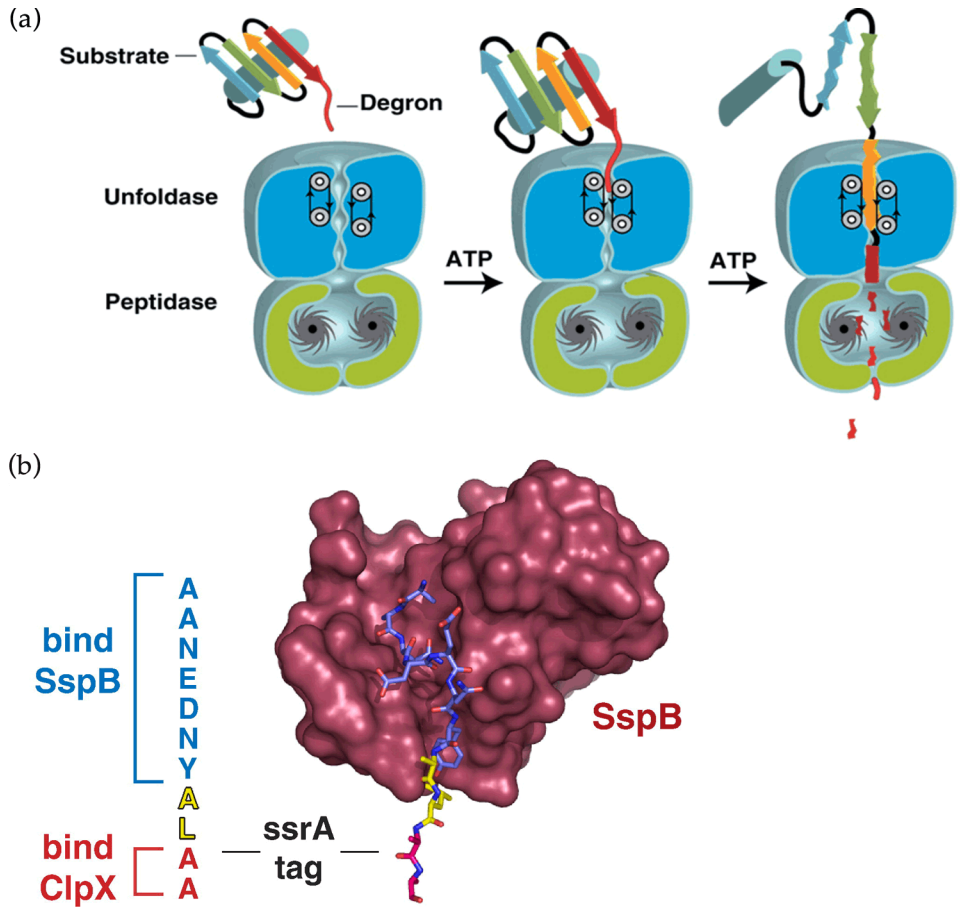


Figure 1.6: ClpXP degradation of tagged proteins is enhanced by binding of the adapter protein SspB to the ssrA tag.
(a) ClpXP proteolytic degradation mechanism [119] and (b) recognition of the C-terminal ssrA tag by SspB and ClpX [120].

When a protein in the cytoplasm of *E. coli* carries the C-terminal ssrA tag, which has the sequence AANDENYALAA, it is marked for degradation by the ClpXP protease. The N-terminal AANDENY portion of the ssrA tag is bound by a groove in one of the subunits of the homodimeric adapter SspB. Both C-terminal tails of the SspB dimer bind to the ClpX pore and thereby aid in the delivery of the substrate tagged for degradation. SspB thus enhances recognition of ssrA-tagged substrates by ClpX and at the same time blocks recognition by ClpA, which also degrades proteins carrying a free C-terminal ssrA tag. The ClpA ATPase can also associate with the peptidase ClpP forming a different AAA+ protease ClpAP.

The two C-terminal alanines and the negatively charged α-carboxylate of the ssrA tag are recognized by the ClpX pore directly.

The separate binary interactions that stabilize the ternary complex of the ssrA-tagged substrate, SspB and ClpX are relatively weak. The ssrA tag binds to the ClpX pore with an affinity of ~1 µM and each SspB tail binds to an N-terminal domain of ClpX with ~20 µM affinity. However, combination of all three contacts results in very strong binding of the ssrA-SspB complex to the ClpX pore with ~15 nm affinity [121,122].

In contrast to the high efficiency of the ssrA/ClpXP degradation machinery proteins targeted for Tat translocation have a substantial retention time in the cytoplasm prior to transport, and the Tat pathway in general has a lower translocation rate than the Sec pathway [107,108]. Therefore, a system using GFP-ssrA as reporter protein may be too fast in the elimination of the reporter and the protein of interest, compared to a possibly slow folding of the POI plus Tat targeting and proofreading.

The ClpXP-dependent degradation rate requires to be balanced for removal of the cytoplasmic fraction of Tat-targeted constructs, while allowing the translocation of slowly folding proteins carrying the degradation tag. This can be achieved by modification or removal of components of the degradation machinery (e.g. SspB, ClpX, ClpP) or by adapting their recognition sequences in the degradation tag [123].

1.6 Aim of the thesis

The motivation for this work stems from the limited understanding of the mechanisms governing protein folding and the observation that nature only uses a comparatively small number of protein folds. Most likely, the largest portion of protein sequence space has not been sampled yet and many more stable protein folds are feasible.

We aimed to search for truly novel, stably folded proteins by using libraries covering unexplored sequence space in combination with a selection system that may be able to target protein folding directly. Such a selection system could be achievable in the form of the bacterial Tat pathway with its folding quality control. To achieve a good correlation between Tat-dependent translocation and the folding properties of the proteins targeted to the Tat pore, we sought to establish suitable reporter systems and fine-tune their parameters to obtain a high dynamic range and minimize escape mechanisms.

To use such an in-vivo selection system targeting protein folding in the search for truly novel proteins, libraries are needed, which encode proteins without significant homology to natural proteins or which essentially cover unexplored sequence space. We aimed to use a previously established secondary structure library, constructed from combinatorial arrangements of binary patterned modules. Additionally, we wanted to seamlessly assemble a fully random, entirely unbiased library of high quality.

We sought to build all libraries for Tat-dependent selections to encode for about 100 amino acids and include a hydrophobic patch of five residues, which should impede bypassing the Tat folding proofreading and may function as nucleation center in the formation of a well packed, solution-shielded hydrophobic protein-core.

Finally, by combining our selection system that targets folding with the libraries covering unexplored sequence space we wanted to test the potential of these tools in the search for truly novel proteins.

2 Results

2.1 *E. coli* translocation systems and selections for folding

We sought to characterize the Tat pathway in comparison to the two major translocation pathways in *E. coli* targeting the Sec translocon to evaluate if and under which circumstances the Tat pathway would be suited to detect the folding state of proteins in order to use it as tool for selections of well-folded proteins.

Using a collection of proteins with diverse folding characteristics, we studied the potential of the reporter β-lactamase for discriminating folding states beyond the published capability to distinguish proteins based on their solubility when targeting the Tat pathway.

Furthermore, we intended to establish a more Tat-exclusive reporter system without survival pressure in selections. Translocation of GFP to the periplasm, in a fluorescent state, had been described to occur only via the Tat pathway [109,112-114]. We pursued to establish GFP as Tat-exclusive reporter by carefully balancing cytoplasmic degradation rate with the time needed for folding quality control and Tat-dependent translocation.

2.1.1 Design of the initial reporter system

To characterize the bacterial transport pathways especially with regard to their suitability for selections for well folded proteins we used the following setup: A mid-to-high copy plasmid (ColE1 Ext2 origin) with an antibiotic resistance marker (Cm^R) different to potential reporter proteins, as e.g. β-lactamase, and a modular design of a Lac-promotor inducible secreted POI-reporter fusion and a LacIQ repressor on the plasmid for tight expression control. The prototype plasmid was constructed from the template pDST22 [124], which is a derivative of the vector pMorph7 [125]. See **Figure 6.2** for a vector map of the main constructs used in this thesis.

A short FLAG M1 tag is present directly after the N-terminal signal sequence. This tag may be used to detect transported proteins, where the signal sequence has been cleaved off leaving the DYDK motif at the very N-terminus of the processed protein.

The restriction sites were chosen by making a list of all sequences of the POIs used for the initial characterization. This sequence list was searched for robust (>95% for ligation and re-cutting) restriction endonucleases that do not cut in the sequences encoding the POIs or in the vector backbone. The recognition sequences for the restriction enzymes should also not encode any unfavorable amino acids for a peptide linker in the used reading frame. **Figure 2.1** shows a schematic layout of the construct design.

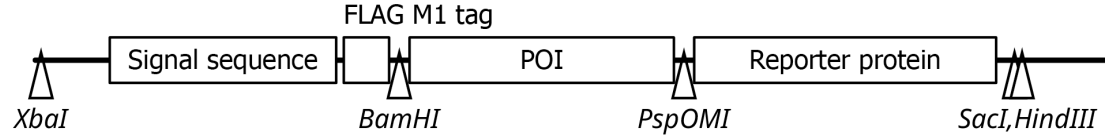


Figure 2.1: Layout of reporter system with employed restriction sites.

For the linker between the POI and the reporter the restriction enzyme PspOMI was chosen. The palindromic recognition sequence (GGGCCC) of PspOMI encodes a glycine followed by a proline in the chosen reading frame. To assess the suitability of the Gly-Pro in the linker sequence a search in the PDB was conducted for the full linker (SGPSG) yielding results of structures containing this sequence in loop or turn conformations, which supported our choice for the restriction site and linker sequence (**Figure 2.2**).

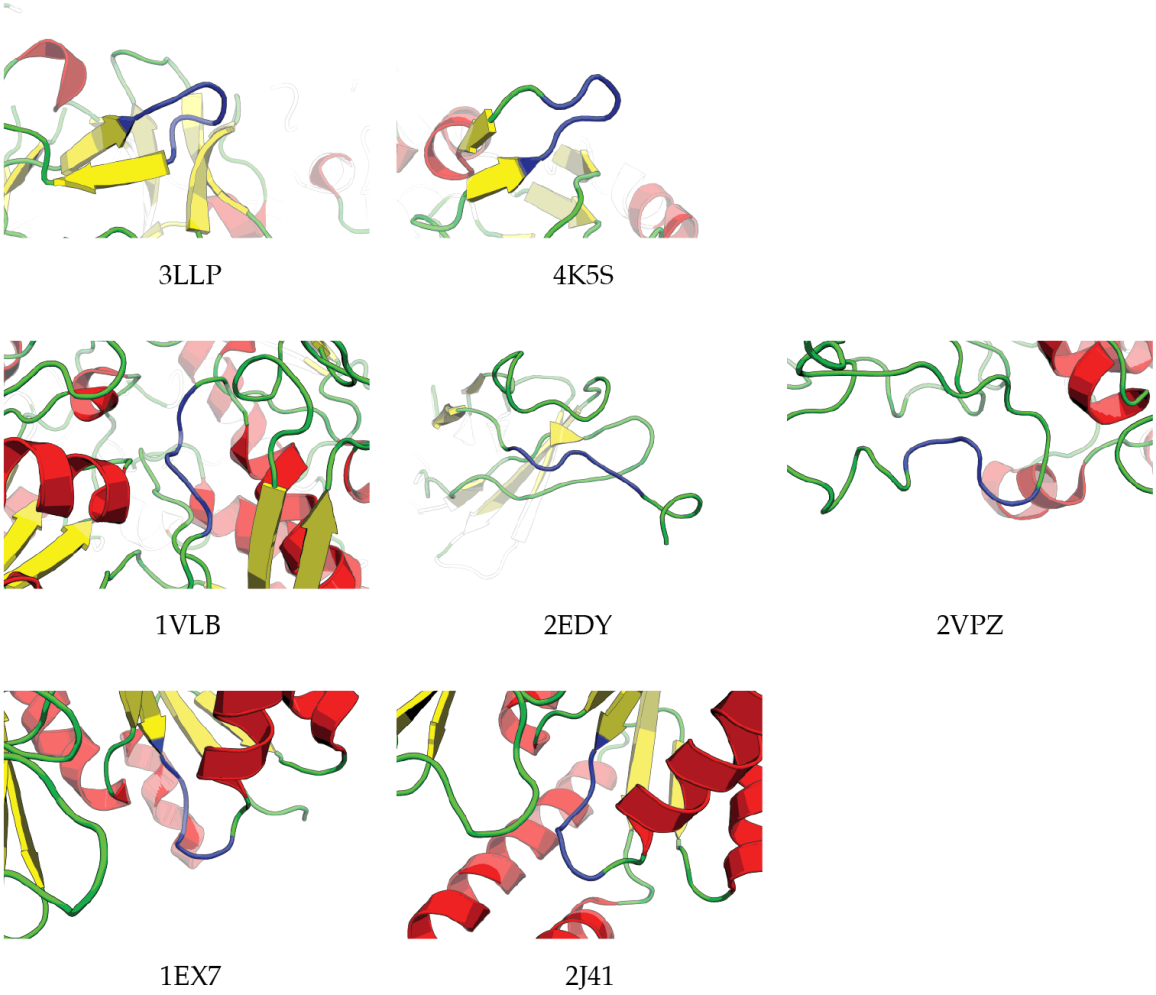


Figure 2.2: Conformations of the peptide linker sequence SGPSG found in PDB structures. The linker, shown in blue color, adopts extended conformations, both in burried and solvent accessible environment; Protein Data Bank accession codes of the molecules below each example.

2.1.2 Four signal sequences targeting 3 bacterial translocation pathways

To target translocation to the periplasm via different bacterial transport pathways we used one signal sequence for the SecB pathway (phoA_ss), one signal sequence for the SRP pathway (DsbA_ss), and two signal sequences for the Tat pathway (TorA_ss and SufI_ss).

The *E. coli* alkaline phosphatase (Swiss-Prot: P00634) PhoA has the signal sequence PhoA_ss (1-21) MKQSTIALALLPLLFTPVTKA. PhoA_ss targets post-translational export via the SecB pathway in an unfolded conformation [126-129].

The *E. coli* Thiol:disulfide interchange protein (Swiss-Prot: P0AEG4) DsbA has the signal sequence DsbA_ss (1-19) MKKIWLALAGLVLAFSASA. DsbA_ss targets co-translational export via the SRP pathway in an unfolded conformation [130].

The *E. coli* Trimethylamine-N-oxide (TMAO) reductase 1 (Swiss-Prot: P33225) TorA has the signal sequence TorA_ss' (1-39). Here, three additional amino acids of the TorA pre-protein were included, as they ensure efficient translocation and processing of the signal peptide [100,107,118,131]. TorA_ss (1-42): MNNNDLFQA**SRRRFLA**QLGGLTVAGMLGPSLLTPRRATA'AQA.

The *E. coli* cell division protein FtsP (Swiss-Prot: P26648), also known as SufI, has the signal sequence SufI_ss (1-27): MSL**SRRQFIQ**ASGIALCAGAVPLKASA.

The Tat signal sequences (consensus motif in bold) TorA_ss and SufI_ss target post-translational export via the Tat pathway in a putatively folded conformation, probably including a folding quality check prior to translocation [104,105]..

For the Tat pathway we sought to evaluate the selection potential for folding characteristics beyond discrimination for protein solubility, which was published for beta-lactamase as reporter [107].

Therefore, we used a set of model proteins having a variety of folding characteristics.

2.1.3 Proteins with different folding behaviors used for characterization

The protein set used to characterize the *E. coli* translocation systems (SecB, SRP, and Tat) is composed of a variety of natural and synthetic proteins. This set spans a wide range regarding folding and solubility of the proteins in order to test especially the folding quality check of the Tat system.

The following proteins were included in the set: the two natural proteins TrxA, Thioredoxin-1 of *Escherichia coli* (Swiss-Prot: P0AA25), and gpD, a N-terminally truncated head decoration protein (also known as major capsid protein D) of the enterobacteria phage lambda (Swiss-Prot: P03712), the designed Ankyrin repeat protein (DARPin) E3_5 (PDB: 1MJ0) [132,133], four designed armadillo repeat proteins Y2CA, Y2MA, Y4CA, and Y4MA [134,135], three proteins from the prototype Secondary Structure Library (SSL) 2ex10 (α10α5βt), 3ex24 (α10βt), and 4ex24 (α10α5t) [53], and two proteins of the second generation SSL2.1 AE564 (R.4.4), AE73 (6.1/3T.D9) [54].

Table 2.1: Set of proteins with different folding and solubility properties used for characterization.

POI	reference	length/aa	solubility	folding	n-mer
TrxA	SP: P0AA25	108	⊕⊕⊕	⊕⊕⊕	mono
gpD	SP: P03712	90	⊕⊕⊕	⊕⊕⊕	mono
E3_5	PDB: 1MJ0	154	⊕⊕⊕	⊕⊕⊕	mono
Y2CA	[134,135]	157	⊕	○	mono/di
Y2MA	[134,135]	157	n/d	n/d	n/d
Y4CA	[134,135]	241	⊕	⊕	mono/multi
Y4MA	[134,135]	241	⊕⊕	⊕⊕	mono
2ex10	[53]	128	⊖	⊖	mono
3ex24	[53]	91	⊖⊖⊖	⊖⊖⊖	multi
4ex24	[53]	100	○	⊖	mono/di/tri
AE73	[54]	47	⊕	○	mono
AE564	[54]	123	○	⊖⊖	di/multi

POI: protein of interest; SP: UniProtKB/Swiss-Prot accession code; PDB: Protein Data Bank accession code; solubility: protein solubility under physiological conditions without fusion to other proteins; folding: tendency of the protein to adapt a stably folded tertiary structure, to show significant signal of secondary structure formation, or not to present exposed hydrophobic patches indicated by low ANS binding; n-mer: oligomeric state(s) of the protein under physiological conditions, assessed mainly by size exclusion chromatography; n/d: not determined.

Figure 2.3: Module composition of POIs from the SSL used for characterization.

2.1.4 Reporters for export: T267A TEM-116 Bla, SF·GFP, and S65T wtGFP

Quantification of export rates needs a reporter whose concentration in the periplasm correlates with an observable phenotype. One such phenotype can be resistance to β-lactam antibiotics. A suitable reporter protein would be β-lactamase, which needs to be present outside the cytoplasm to confer resistance. The β-lactamase used in this work has one amino acid exchange Thr 267 → Ala (T267A) compared to the TEM-116 β-lactamase (UniProtKB: Q79DR3).

Such a system of resistance to β-lactam antibiotics has been employed for in-frame selections [136,137] as well as selection for solubility of proteins using the Tat pathway [107]. **Figure 2.4** shows the general principle of a Tat-dependent selection system using β-lactamase as reporter.

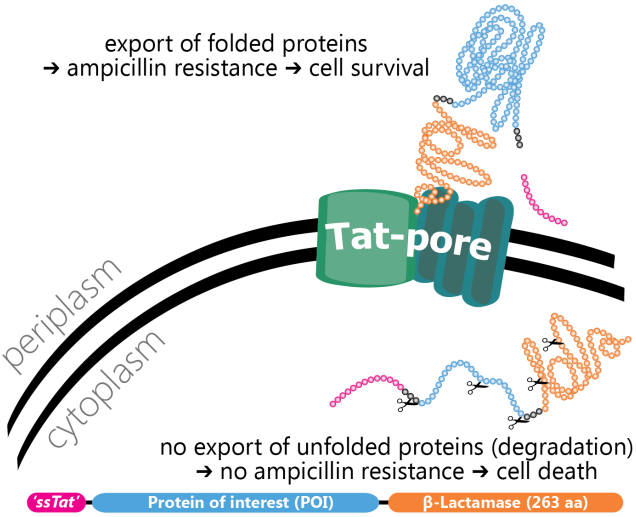


Figure 2.4: Scheme of the selection system based on β-lactamase (Bla setup) targeting the Tat pathway.

Another suitable phenotype would be periplasmic fluorescence. GFP was described to be fluorescent in the periplasm only when exported via the Tat pathway [109,112-114], which makes it an excellent reporter for characterizing the transport via the Tat pore, while excluding functional transport via other pathways. In this work, we used superfolder GFP (SF·GFP) [117] as well as the S65T mutant [111,138] of *A. victoria* GFP.

GFP has to be fully folded in the cytoplasm before export via the Tat pathway can occur. This can lead to a high fluorescence signal that is located in the cytoplasm, which would no longer correlate with export rates. One possibility to restrain cytoplasmic fluorescence is to include a degradation tag (see 2.1.8), which ensures that cytoplasmic GFP is rapidly cleared.

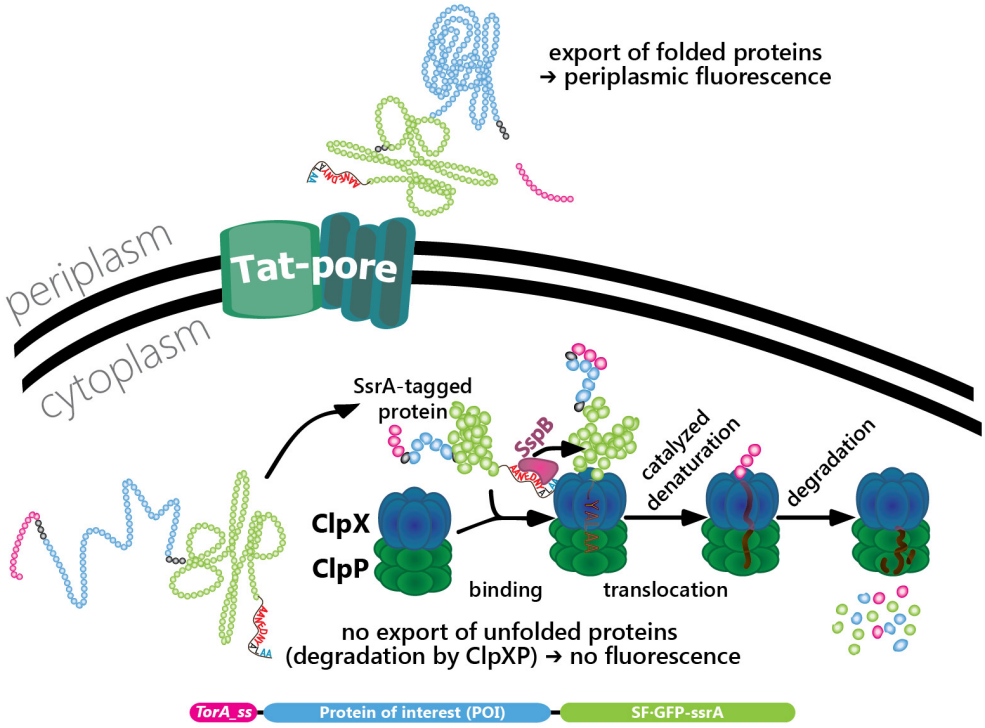


Figure 2.5: GFP setup with a C-terminal degradation tag used in Tat-dependent selections for periplasmic fluorescence.

2.1.5 Bla assays in liquid culture

The export of a set of proteins with different folding properties was first characterized in the Bla setup. Liquid cultures were prepared in small scale and growth under different conditions was monitored by measuring a time series of the absorption at 595 nm ($Abs_{595} = OD_{600}/2.4$ for 200 µl and the plate types used here). The *E. coli* Top10F' cells were shaken at 37°C in 96-well plates having 200 µl 2YT per well with final concentrations of 25 µg/ml Cm, 0.5 mM IPTG, and 20 µg/ml or 100 µg/ml Amp. Control plates without ampicillin (Amp) and without both IPTG and Amp were recorded in parallel. On the control plates, growth curves for all constructs were almost identical.

Yet, under selective conditions we observed little reproducibility of the recoded growth curves in liquid cultures, due to a strong influence of plate type and shaking conditions. These fluctuations were most pronounced for constructs targeted to the Tat pathway via TorA_ss or SufI_ss (**Figure 2.6**).

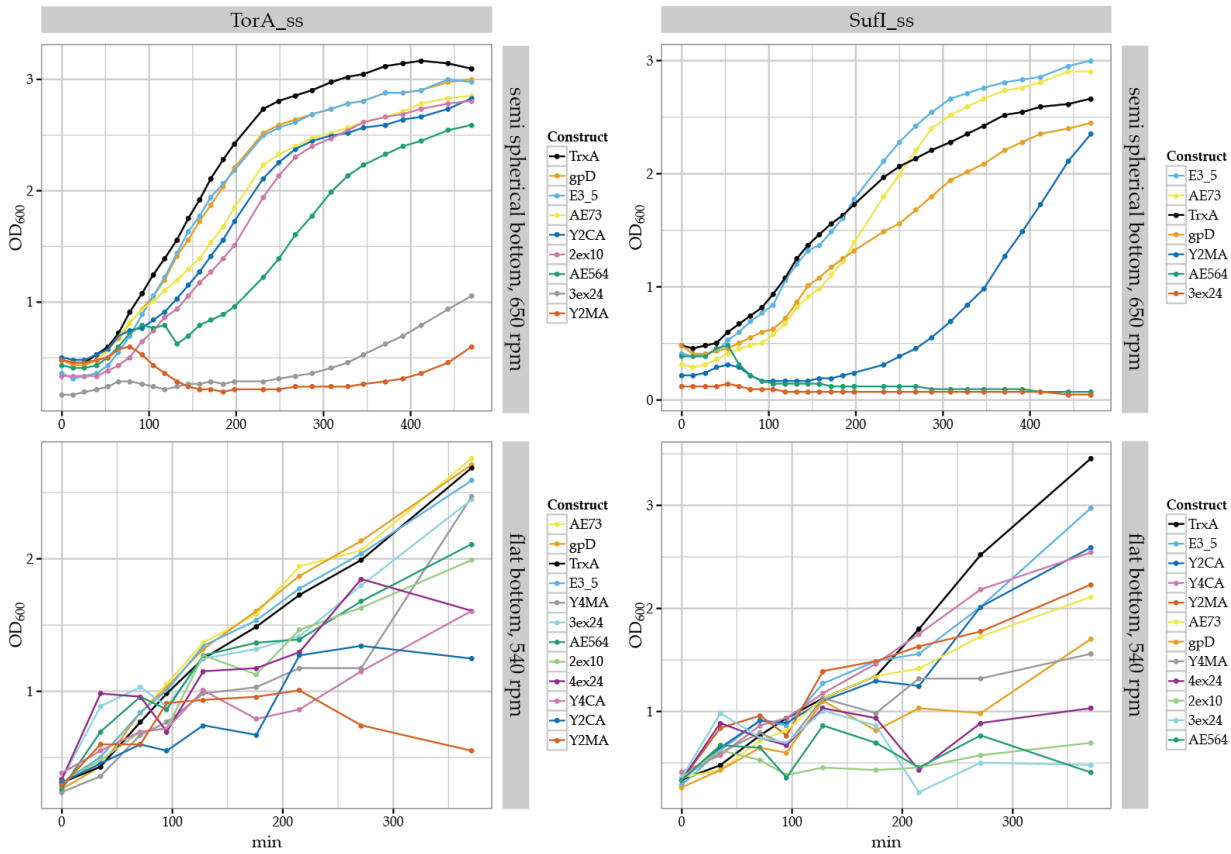


Figure 2.6: Growth curves for liquid culture Bla assays showed little reproducibility for TorA_ss and SufI_ss constructs.
Similar selective conditions were used, variations could be attributed to shaking conditions and dead cells blocking the optical path (see **Figure 2.7**).

The liquid culture assays showed a certain reproducibility when using identical parameters, but minor changes in plate type or shaking conditions had a strong influence on the recorded growth curves, especially for Tat-targeted constructs. The reason for this unexpected behavior was found in the presence of large amounts of dead cells due to the high survival pressure of this setup.

Absorbance readings were not representative of the cell growth and proliferation, as dead cells sank to the bottom of the plate and falsified the measurements (**Figure 2.7**).

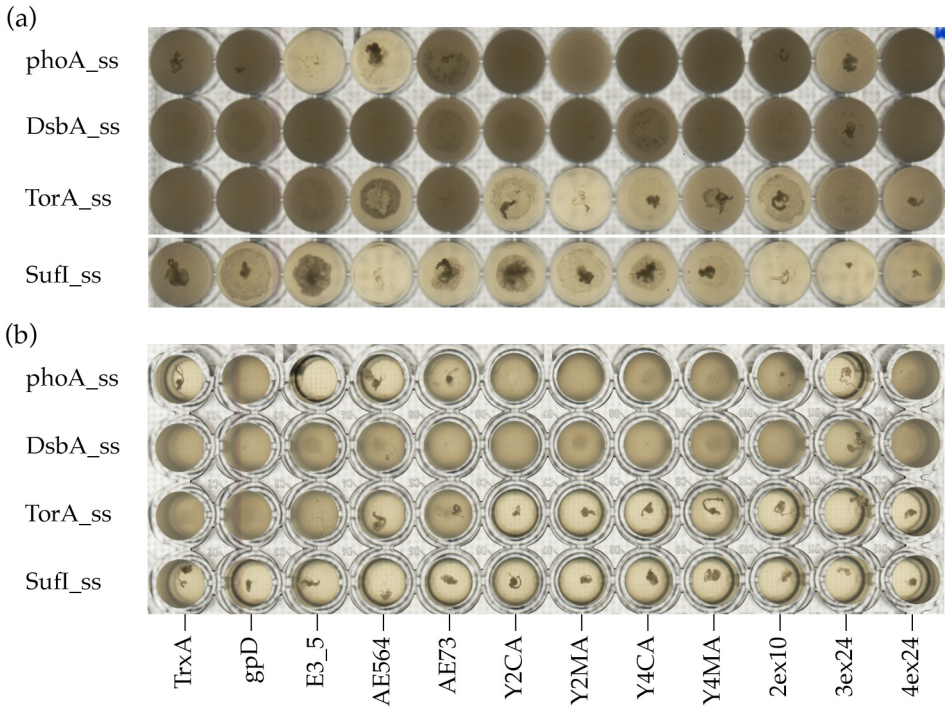


Figure 2.7: Influence of 96-well plate type and shaking condition on Bla assay in liquid culture. a) Nunc semi spherical bottom, shaking at 650 rpm; b) Nunc MaxiSorp - flat bottom, shaking at 540 rpm. The settlement of dead cells can be seen as darker spots or rings, absorbance readings do no longer correlate with cell proliferation.

2.1.6 Bla assay on solid media plates, droplets of dilution steps

Due to the limitations of the Bla assays in liquid culture, further Bla-dependent export assays were performed on solid media plates by spotting a dilution series of droplets. These solid media Bla assays showed better reproducibility than the assays in liquid medium. For selective conditions two different stringencies were initially applied, low stringency plates containing LB-agar with 25 µg/ml Cm, 0.5 mM IPTG, and 20 µg/ml Amp, and high stringency LB-agar plates with 25 µg/ml Cm, 1 mM IPTG, and 100 µg/ml Amp. **Figure 2.8** shows plots indicating the last dilution step where growth of more than ten single colonies could be observed for low stringency (a) and high stringency (b) as well as scans of the high stringency plates overlaid (c).

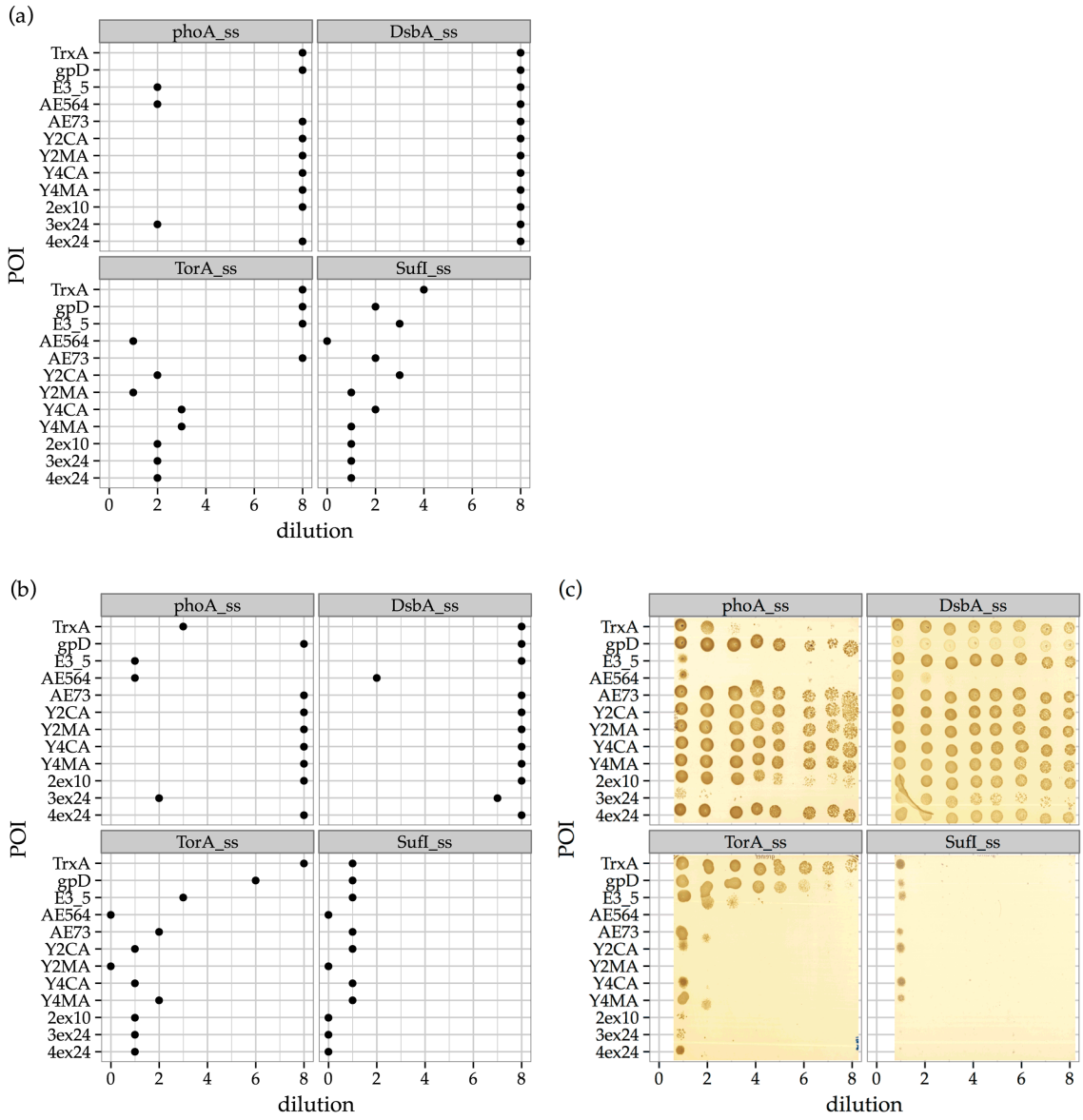


Figure 2.8: Bla assay on solid media plates of the 12 POIs with 4 signal sequences at 37°C. Droplets of 1:8 dilution series were printed on selective solid media LB-agar plates having low stringency (a) with 0.5 mM IPTG and 20 µg/ml Amp, or high stringency (b) with 1 mM IPTG and 100 µg/ml Amp. The dots indicate the last dilution step where growth of more than 10 single colonies could be observed. For comparison, scans of the colonies on high stringency plates are shown in (c).

A ranking of the POIs was compiled, according to their transport efficiencies using the four signal sequences targeting the respective translocation pathways (phoA_ss → SecB, DsbA_ss → SRP, TorA_ss and SufI_ss → Tat) in *E. coli* Top10F'.

Table 2.2: Ranking of export rates of signal-sequence–POI–Bla constructs on selective solid media plates.

phoA_ss	DsbA_ss	TorA_ss	SufI_ss
1: Y4CA	1: gpD	1: TrxA	1: TrxA
1: Y2CA	1: TrxA	2: gpD	2: E3_5
1: Y4MA	1: 4ex24	3: E3_5	2: Y2CA
1: Y2MA	1: E3_5	4: Y4MA	3: gpD
1: gpD	1: AE73	5: AE73	4: AE73
1: AE73	1: Y4MA	6: Y4CA	4: Y4CA
1: 4ex24	1: 2ex10	6: 4ex24	5: Y2MA
2: 2ex10	1: Y2CA	6: Y2CA	5: Y4MA
3: TrxA	1: Y2MA	7: 3ex24	6: 4ex24
4: E3_5	1: Y4CA	8: 2ex10	6: 2ex10
4: AE564	2: 3ex24	9: AE564	7: 3ex24
5: 3ex24	3: AE564	9: Y2MA	8: AE564

The highest ranks (lowest numbers) signify that cells carrying these constructs grew at higher dilutions or ampicillin concentrations, compared to the other tested constructs with the same signal sequence. Constructs with the same rank were not distinguishable in this assay.

Export rates of POI-Bla fusions using the signal sequence SufI_ss targeting the Tat pathway were considerably lower than for TorA_ss. Only with low ampicillin concentrations of 20 µg/ml colonies could be observed at higher dilution steps (**Figure 2.8a**).

Expression of fusion proteins was checked for TorA_ss and SufI_ss constructs by anti-Bla western blots of whole cells lysates after 4 h expression with 1 mM IPTG at 37°C (**Figure 2.9**). Almost all proteins were expressed above detectable levels, albeit constructs with low solubility and unfavorable folding properties showed much weaker bands.

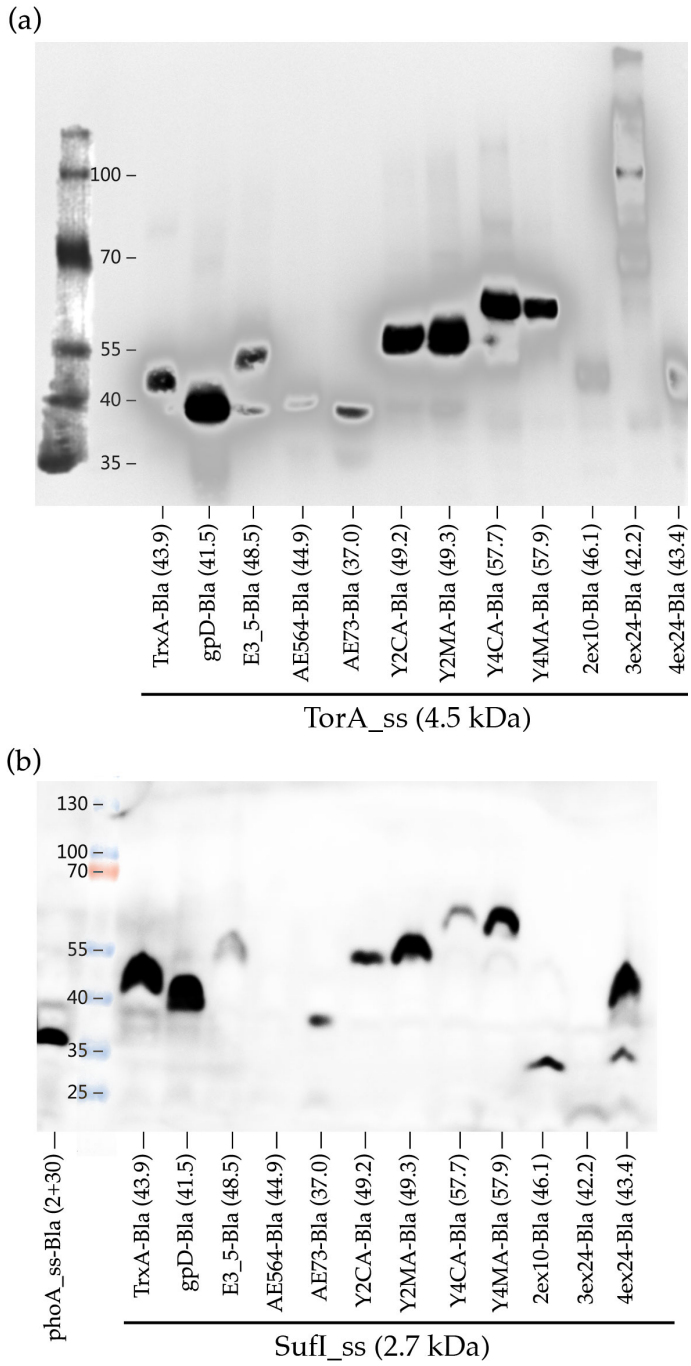


Figure 2.9: Anti β-lactamase western blots for (a) TorA_ss and (b) SufI_ss constructs. Mass in kDa of fusion proteins without signal sequence given in parentheses.

2.1.7 Translocation of SF·GFP to the periplasm via different pathways

Transport of fluorescent GFP to the periplasm of *E. coli* had been reported only via the Tat pathway.

The four signal sequences (phoA_ss → SecB, DsbA_ss → SRP, TorA_ss and SufI_ss → Tat) were fused directly to superfolder GFP. The constructs were expressed in *E. coli* Top10F' with 1 mM IPTG at 32°C for 16 h using 5 ml 2YT with 25 µg/ml Cm.

Fluorescence of whole cells, the periplasmic fraction, and B-per II lysed spheroblasts was measured. It was observed that, unlike other tested variants of GFP, SF·GFP is able to fold properly in the periplasm when exported via SRP. The periplasmic fraction of DsbA_ss-SF·GFP showed a substantial fluorescence signal, which had half the intensity of the periplasmic fraction of TorA_ss-SF·GFP translocated via the Tat pathway (**Figure 2.10**: Periplasm).

It was later reported independently that SF·GFP is fluorescent in the periplasm when translocated via the SRP pathway [139].

The construct SufI_ss-SF·GFP showed no significant fluorescence over background in any fraction. The low fluorescence for SufI_ss fusion proteins and their inefficient Tat-dependent translocation was observed for SufI_ss–POI–SF·GFP constructs as well.

Compared to the other signal sequences, the Tat-dependent signal sequence TorA_ss lead to the highest fluorescence signal of SF·GFP in the periplasm (**Figure 2.10** Periplasm). However, only a small percentage of the expressed TorA_ss–SF·GFP was exported. And although the fluorescence intensities of the different cellular fractions were not directly comparable, due to different quenching effects and buffer conditions, the larger portion of TorA_ss–SF·GFP seemed to be located in the spheroblasts (**Figure 2.10**).

Therefore, to be able to correlate whole-cell fluorescence with Tat-dependent translocation the large non-periplasmic GFP signal has to be eliminated.

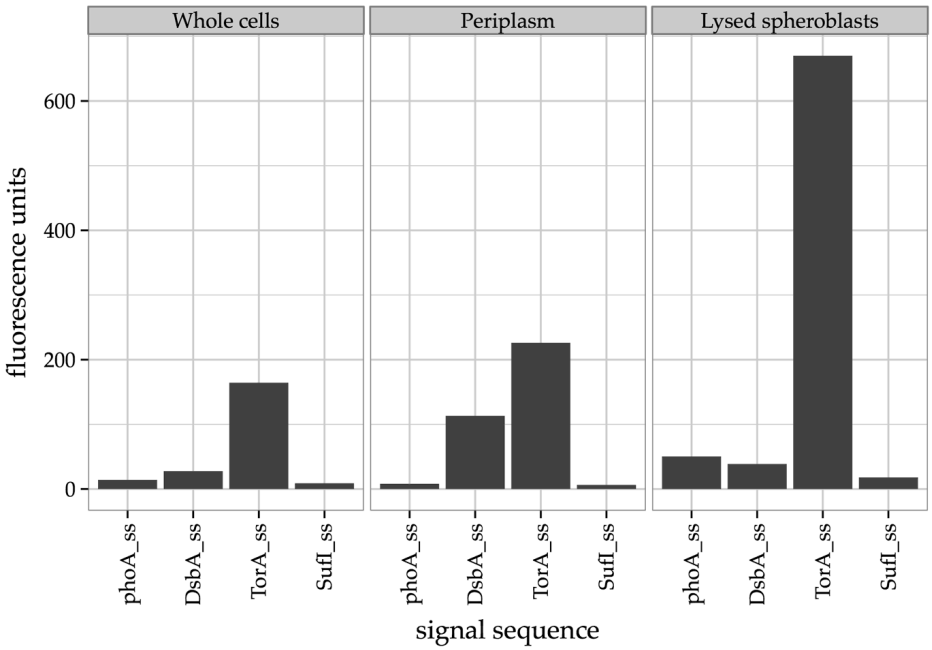


Figure 2.10: SF·GFP fluorescence readings in whole cells and after fractionation.
Measurements are plotted with a common scale on the y-axis, although fractions are not directly comparable due to different buffer conditions and quenching effects.

2.1.8 GFP-ssrA and strains impaired in SspB, ClpXP degradation

To efficiently eliminate the cytoplasmic fraction of GFP, we fused the ssrA degradation tag (amino acid sequence: AANDENYALAA) C-terminally to SF·GFP, which ensures a fast cytoplasmic elimination of tagged proteins by ClpXP.

First, we tested fusion proteins without degradation tag to estimate if the potent ssrA degradation tag would still allow sufficient Tat-dependent translocation.

A set of proteins carrying one of the Tat-dependent signal sequences, SufI_ss or TorA_ss, was expressed as SF·GFP fusion without degradation tag and the fluorescence of the periplasmic fraction was measured (**Figure 2.11**). SufI_ss-SF·GFP constructs showed very weak fluorescence readings. For this reason, the signal sequence SufI_ss was not used in further experiments with the GFP system.

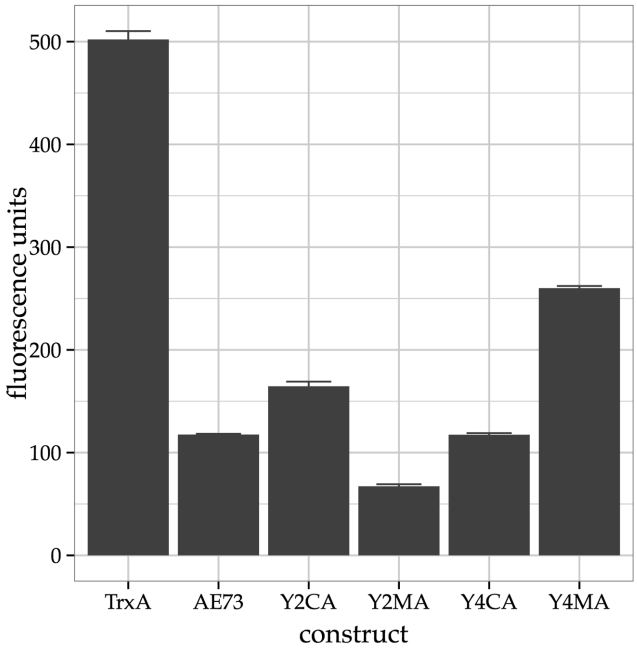


Figure 2.11: Periplasmic SF·GFP fluorescence (without degradation tag) of fusion proteins with TorA_ss.

Having measured the periplasmic fraction of unconstrained Tat-dependent translocation of the GFP constructs without ssrA degradation tag, we tried to quantify how much the ClpXP-dependent degradation would reduce export rates. The GFP fluorescence of constructs additionally carrying the C-terminal ssrA degradation tag was determined for whole cells and the periplasmic fraction.

In strains with an intact SspB and ClpAP/XP degradation machinery (as in Top10F', the strain employed here), the degradation of ssrA-tagged proteins was very fast. Only the positive control TorA_ss–TrxA–SF·GFP-ssrA was translocated via the Tat pathway to the periplasm in detectable quantities. Its fluorescence signal was located entirely in the periplasm. However, the signal intensity was low and close to background fluorescence, especially compared to the periplasmic fraction of the construct TorA_ss–TrxA–SF·GFP without degradation tag.

To increase the dynamic range of the SF·GFP-ssrA system and obtain similar fluorescence readings as for the periplasmic fraction of the constructs without degradation tag, while still minimizing non-periplasmic GFP, ssrA-tagged constructs were expressed in *E. coli* strains lacking distinct components of the SspB and ClpAP/XP degradation machinery. The constructs were targeted to the Tat pathway via the TorA signal sequence. Expression was performed with 80 µM IPTG at three different temperatures (25°C, 31°C, 37°C) in Δ*sspB*, Δ*clpP*, Δ*clpX*, and Δ*uvrA* as an degradation-unrelated control. All deletion strains were obtained from the Keio collection [140], where they were exposed to similar selection and growth procedures.

At three time points after induction (3 h, 6 h, 9 h) whole-cell fluorescence was determined in a flow cytometer for the TorA_ss–POI–SF·GFP-ssrA constructs (**Figure 2.12**).

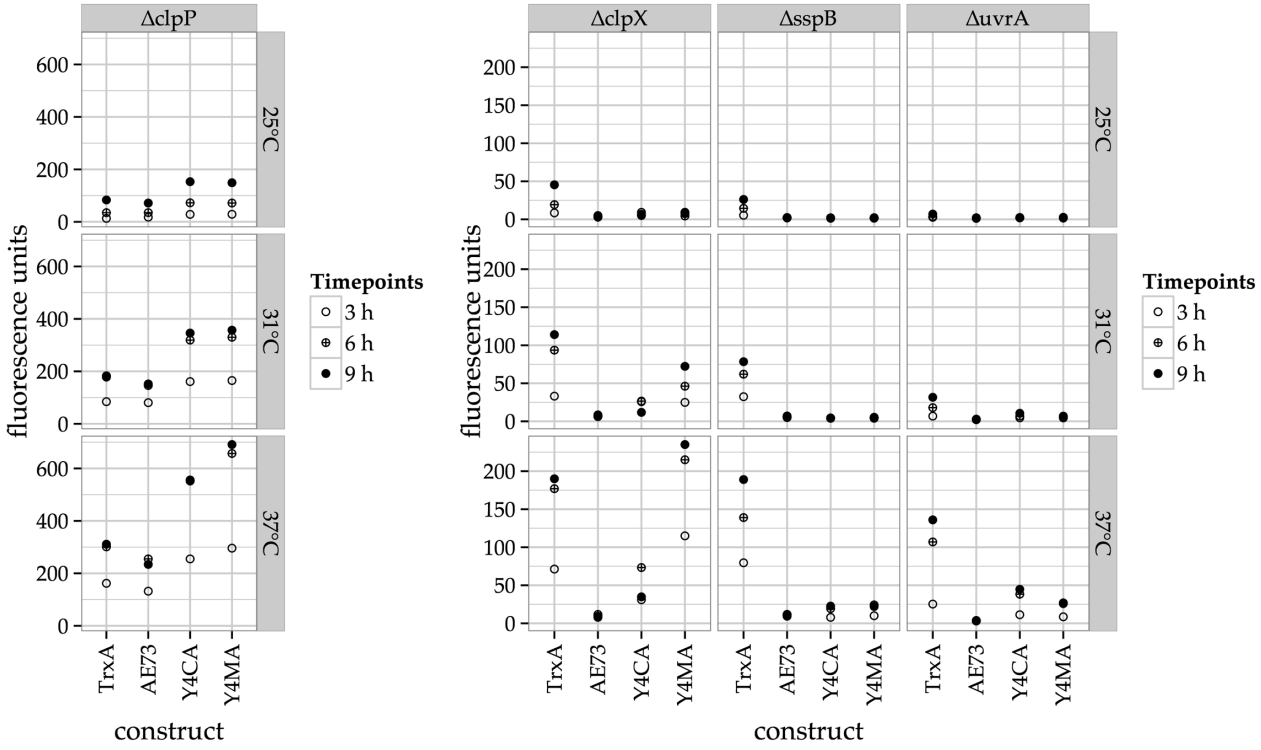


Figure 2.12: Medians of TorA_ss–POI–SF·GFP–ssrA fluorescence measured in FACS for different deletion strains and expression temperatures.

Localization of the GFP fluorescence (**Figure 2.13**) was determined in a plate reader after the last flow cytometry measurement, by cold osmotic shock fractionation.

Removal of the peptidase ClpP in the Δ*clpP* strain showed the highest increase in total cell fluorescence and especially in the cytoplasmic fraction, as constructs that cannot be translocated via the Tat pathway remained intact and fluorescent in the cytoplasm due to the missing of a crucial component of the ClpXP degradation machinery. The reduction of the degradation rate is less pronounced in the Δ*clpX* strain, which is lacking the ClpX unfoldase. In Δ*clpX*, the cytoplasmic fraction did show significant fluorescence over background for the constructs TrxA and the armadillo Y4MA, indicating that the delay of proteolytic degradation was still to high. Removal of the adapter protein SspB in the strain Δ*sspB* lead to cytoplasmic fluorescence levels

close to background, with only the positive control TrxA showing a weak fluorescence signal located in the cytoplasm. Especially at lower temperatures Δ*sspB* showed a higher dynamic range in flow cytometry compared to Δ*uvrA*, which was used as ClpXP unrelated control originating from the same Keio collection as the other deletion strains. The dynamic range of constructs expressed in Δ*sspB* was even more pronounced when compared to the previously used laboratory strain TOP10F'.

For all strains except Δ*clpP*, whole-cell fluorescence increased continuously over time especially for lower temperatures. At 37°C Δ*clpX* showed only a small increase from 6 h to 9 h for TrxA and Y4MA, whereas Y4CA showed highest fluorescence readings after 6 h of expression, possibly due to cytoplasmic aggregation.

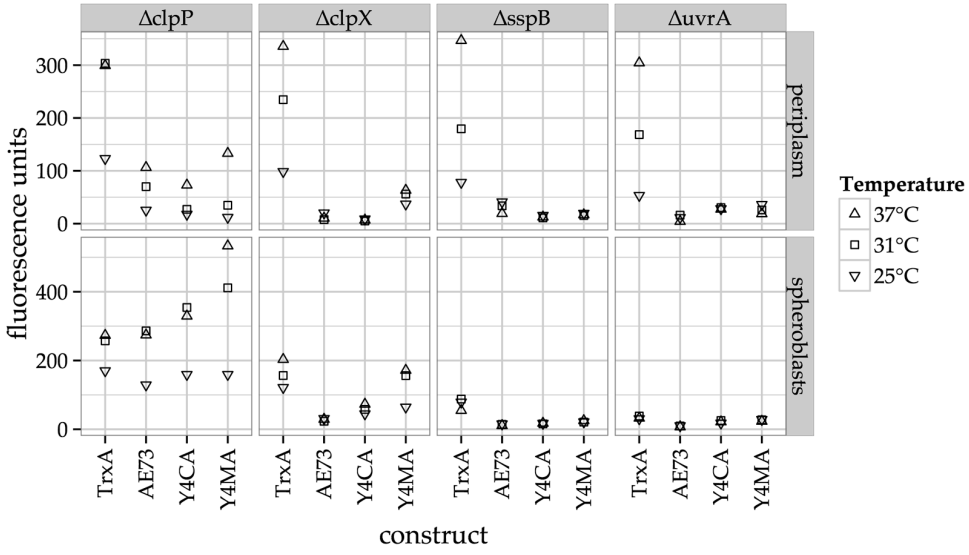


Figure 2.13: Localization of GFP fluorescence for TorA_ss–POI–SF·GFP–ssrA proteins. Constructs were expressed with 80 µM IPTG in four different deletion strains and at three different temperatures. After 9 h expression the cells were lysed by cold-osmotic shock and fluorescence was recorded separately for the periplasmic fraction and the lysed spheroblasts.

As the fluorescence of the tested constructs in Δ*sspB* was located almost exclusively in the periplasm and showed an improved dynamic range, we decided to further investigate the contribution of SspB to cytoplasmic ClpXP-dependent degradation and potential modulations regarding improved Tat-dependent translocation.

2.1.9 Weakened degradation tags as alternative to deletion strains

Many processes in *E. coli* depend on the SspB/ClpXP degradation machinery for the control of cellular protein levels. Instead of perturbing the whole system by removal of a commonly used component like the SspB protein, the degradation of specific ClpXP targets may also be modulated by altering the recognition sequence of SspB in the ssrA tag. The wild-type C-terminal ssrA tag, which we used before, has the sequence AANDENYALAA.

A degradation tag that is not (tightly) bound by SspB but still recognized by ClpXP has been described earlier [123]. The weakened degradation tag was here termed SclpX and has the sequence DDAAAAADLAA. A further variant of the tag, termed Sprc, with the sequence

DDAGVGTDLAA was tested for reduced periplasmic degradation by the tail-specific protease (Tsp/Prc) [141]. We used four test proteins (TrxA, E3_5, AE73, AE564) in the TorA_ss–POI–SF·GFP setup with one of the respective C-terminal degradation tags (ssrA, SclpX, Sprc) for characterization of enhanced Tat-dependent translocation due to reduced cytoplasmic degradation. Expression was performed at 30°C or 37°C with 200 µM IPTG in three different *E. coli* strains (DH5α, Δ*sspB*, Δ*prc*) and whole-cell fluorescence was measured in flow cytometry 3.5 h and 7 h after induction with 200 µM IPTG.

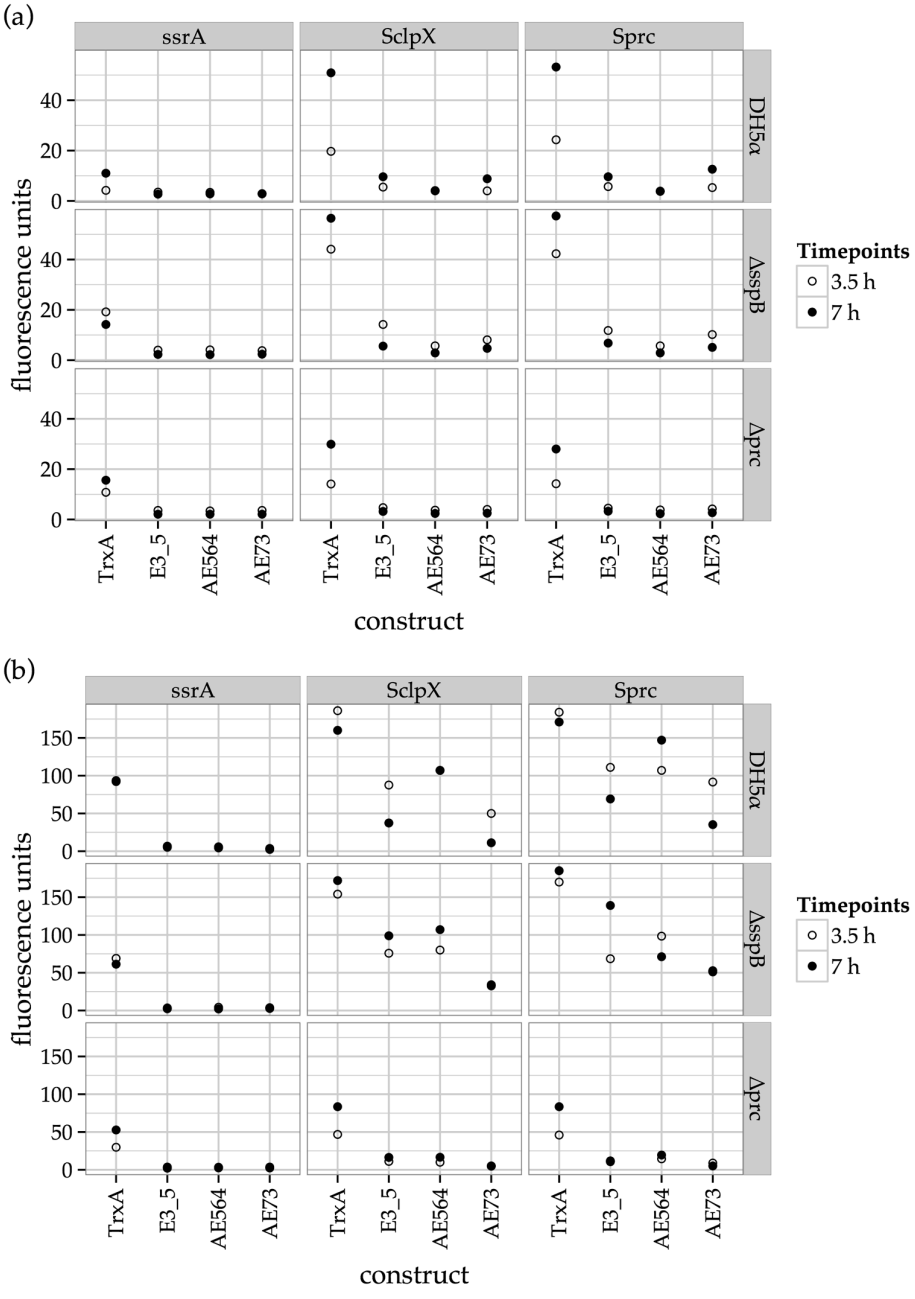


Figure 2.14: Medians of whole-cell fluorescence in FACS for TorA_ss–POI–SF·GFP constructs. Proteins were expressed with 200 µM IPTG carrying one of three different degradation tags (ssrA, SclpX, Sprc) and using one of three *E. coli* strains at (a) 30°C or (b) 37°C.

Constructs carrying the weakened degradation tags SclpX or Sprc showed high whole-cell fluorescence signals also for the negative controls, when expressed in DH5α or Δ*sspB* at 37°C

with 200 µM IPTG (**Figure 2.14**); both for the aggregation prone AE564 and the 47 amino acid short, unstructured but soluble AE73. This was not observed for expressions of SclpX-tagged constructs in DH5α at 37°C with 150 µM IPTG. SsrA-tagged constructs showed little or no increase in fluorescence from 3.5 h to 7 h when expression was induced with 200 µM IPTG; in comparison, 80 µM IPTG resulted in whole-cell fluorescence signals intensifying up to nine hours after induction (**Figure 2.12**).

The use of a weakened degradation tag instead of the SspB deletion strain helped to avoid systemic perturbation of the cellular ClpXP degradation machinery. The C-terminal tags SclpX and Sprc, designed not be recognized by the adapter protein SspB, showed higher fluorescence for the positive control TrxA, both in DH5α and the deletion strains Δ*sspB* and Δ*prc*. Expression of the negative controls with a weakened degradation tag at higher levels, 37°C and 200 µM IPTG, revealed a mechanism of generating false positive signals located in the cytoplasm. This was not due to the modification of the degradation tag, as it could be observed for other constructs even at lower expression levels and with the strong degradation tag (see 2.5.3).

There was almost no increase in fluorescence after 7 h expression at 30°C for SclpX or Sprc-tagged proteins in Δ*sspB*, where SspB is not present. This indicated that the mutated tags SclpX and Sprc were not recognized by SspB and that the SspB:ssrA interaction could be modulated as well by modification of the degradation tag instead of removing SspB.

2.1.10 S65T·GFP as folding reporter

Superfolder GFP is an extremely stable protein and can remain folded and fluorescent even in constructs that aggregate in the cytoplasm. Aggregation may withdraw such constructs from degradation, while the fused SF·GFP remains fluorescent. This leads to the detection of false positives where cell fluorescence is no longer coupled to Tat-dependent translocation to the periplasm.

Therefore, we assessed the suitability of the S65T [138] mutant of *A. victoria* GFP as reporter protein. This variant of GFP is improved in its fluorescence properties but close to the wild-type regarding folding stability and loses its fluorescence upon aggregation [142] [143].

A set of proteins was expressed as TorA_ss–POI–S65T·GFP-SclpX fusions in DH5α and compared to fusion constructs with SF·GFP-ssrA or SF·GFP-SclpX. Whole-cell fluorescence was measured in flow cytometry and after cold osmotic shock preparation the periplasmic fraction and the spheroblasts were measured in a fluorescence plate reader.

The S65T·GFP-SclpX fusions showed very low fluorescence signals, similar to the SF·GFP-ssrA constructs with the strong degradation tag. The positive control TorA_ss–TrxA–S65T·GFP-SclpX had a GFP fluorescence signal that was more than 20 times lower than the respective SF·GFP-SclpX variant.

Furthermore, the false positive constructs from round 2α of the MOAL selection (2.5.2) showed localization of the fluorescence in the cytoplasm even as S65T·GFP-SclpX fusions (**Figure 2.42**).

Considering that the wild-type like S65T·GFP could not be used to eliminate this type of false positives with strong cytoplasmic fluorescence along with the greatly improved fluorescence properties of superfolder GFP, we did not utilize S65T·GFP as reporter protein for selections.

2.1.11 Effects of co-expression of DnaK/J chaperones

TorA_ss–POI–SF·GFP-ssrA constructs expressed in TOP10F' only showed a weak fluorescence signal for the positive control TrxA. One major cause for this is the efficient recognition and degradation of ssrA-tagged proteins, as could be shown in the experiments where the degradation rate had been reduced (see 2.1.8 and 2.1.9).

Additionally, the retention time in the cytoplasm needed for correct folding and proofreading of the Tat translocase may influence export rates and periplasmic fluorescence.

To estimate the contribution of faster folding aided by chaperones on Tat-dependent transport [144], we co-expressed the major chaperone system DnaK/DnaJ with a set of proteins in the TorA_ss–POI–SF·GFP-ssrA format in *E. coli* TOP10F' cells. Fluorescence was measured for whole cells in a flow cytometer.

The DnaK/J plasmid eLIC_031_pAk_lacIq-T5-DnaK_DnaJ-his6 was constructed from eLIC_031 by ligation independent cloning.

Co-expression of DnaK/J chaperones did not result in a higher dynamic range of the fluorescence readings of the tested SF·GFP-ssrA constructs. However, it did result in a faster establishment of the observed profile, where the positive control TrxA was already showing higher fluorescence after 3 h of expression. The fluorescence histograms after 6 h expression were very similar with and without the co-expression of DnaK/J (**Figure 2.15**).

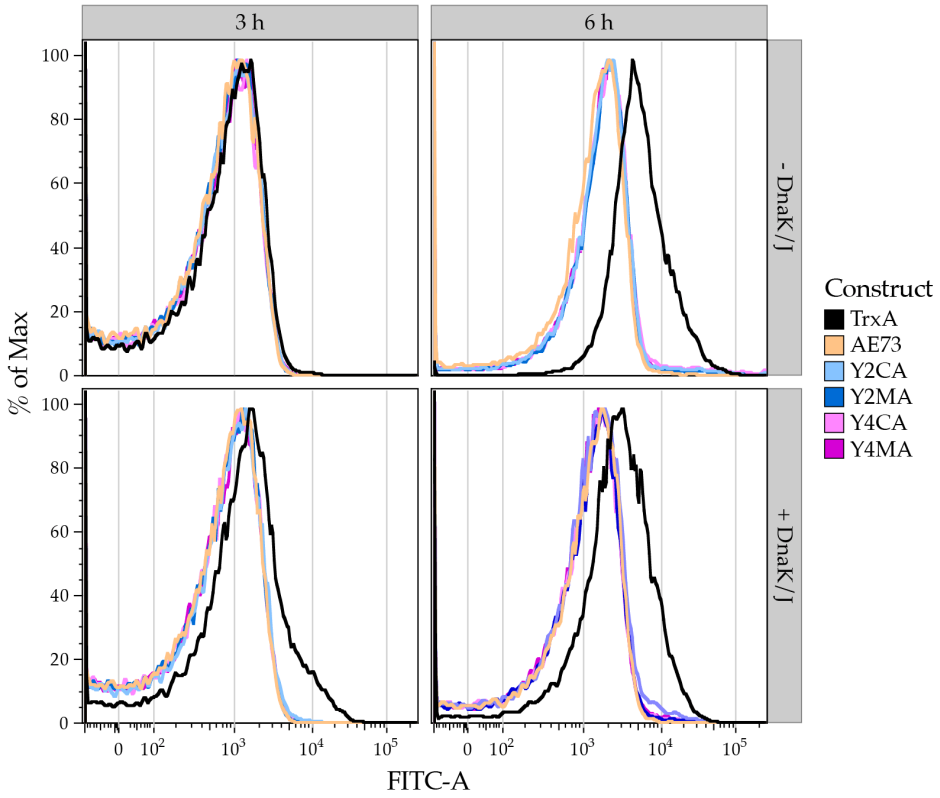


Figure 2.15: FACS profiles of TorA_ss–POI–SF·GFP-ssrA without and with co-expression of DnaK/J chaperones.
Six constructs were expressed in TOP10F' and analyzed after 3 h and 6 h in flow cytometry.

2.1.12 Effect of methionine in peptide linker after the signal sequence

Two different vector formats were created, one for the characterization of the model proteins using the restriction sites BamHI and PspOMI, the other for selections on the previously constructed SSL2.1 using the restriction sites NcoI and BamHI of the library. The restrictions sites for SSL2.1 were given by the design of the finished library SSL2.1 [54]. As NcoI restriction sites were present in the sequences of the POIs, a different restriction endonuclease, PspOMI, was selected for cloning of the POIs, which were used for characterization (2.1.1).

The methionine encoded in the used reading frame of the restriction site of NcoI is located in the linker after the signal sequence. The presence of such an exposed hydrophobic amino acid close to the end of the signal sequence may impede translocation via the Tat pathway. In conjunction with an upstream cryptic promotor region, the methionine might further serve as an unexpected translation initiation site resulting in the expression of proteins without signal peptide.

To examine its effect on Tat-dependent export to the periplasm, we quantified the translocation of TrxA and AE564 with and without the methionine after the signal sequence. The GFP fluorescence of 4L_TorA_ss–TrxA–SF·GFP-ssrA (with Met in linker) and TorA_ss–TrxA–SF·GFP-ssrA (without Met in linker) were measured for Δ*sspB* whole cells in a flow cytometer and for the periplasmic fraction in a fluorescence plate reader. The construct with the methionine after the signal sequence showed only 42% of the fluorescence (median of GFP signal) in FACS

compared to the linker type without Methionine, the percentage after periplasmic extraction was 56% (**Figure 2.16**).

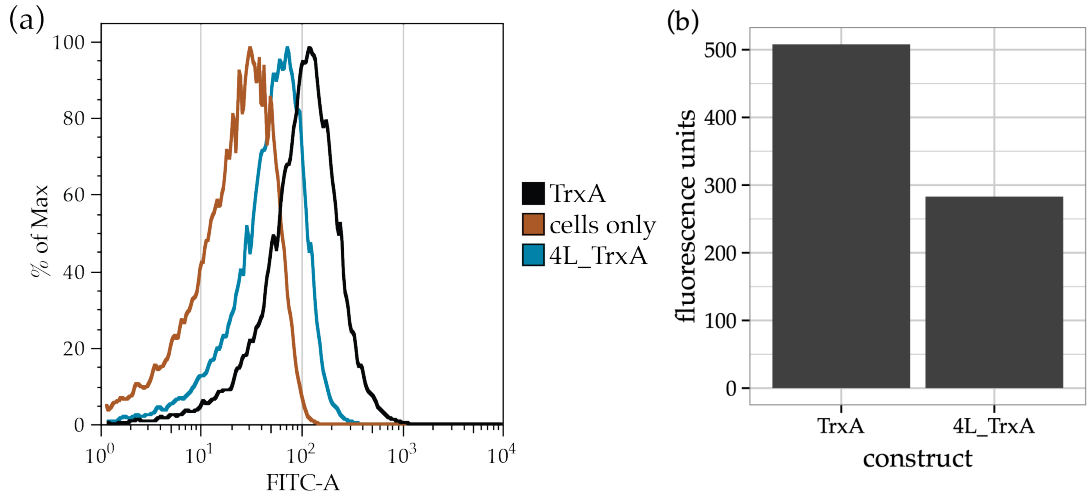


Figure 2.16: GFP fluorescence signals for TrxA without methionine and 4L_TrxA with methionine after the TorA_ss.
(a) Whole cells measured in flow cytometry, and (b) periplasmic signal after cold-osmotic shock fractionation.

The export rates of TrxA and AE564 with both linker types were also determined in the Bla setup, where TorA_ss–TrxA–Bla showed good export rates irrespective of the presence of the methionine in the linker. AE564-Bla showed significantly better export rates without methionine after the signal sequence, as colonies were present three (1:8) dilution steps beyond 4L_TorA_ss–AE564–Bla (with Met in linker) on plates with 25 µg/ml carbenicillin incubated at 30°C (**Figure 2.38**).

The ATG codon, which is part of the NcoI restriction site and encodes the methionine in the linker could either be recognized as additional start codon or the exposed hydrophobic methionine in the linker could hinder translocation. In both cases the amount of translocated protein would be decreased. An additional start codon would lead to the expression of two variants, one with the TorA_ss targeting the protein for Tat-dependent translocation and the second variant without signal sequence for export. This could reduce the amount of protein targeted to the Tat pore compared to the linker sequence without additional ATG codon. Irrespective of the exact mechanism, the linker format without ATG codon proved to be better suited for our Tat-dependent reporter setups.

2.2 Libraries

The initial library used for selections towards well-folded proteins was the Secondary Structure Library (SSL) based on binary patterning. It was designed and constructed by T. Matsuura and A. Ernst [53]. The revised version SSL2.1, employed here, had been improved regarding the distribution of isoelectric points of the encoded proteins and was built with a higher experimental diversity [54].

In a first step, we attempted to resolve the very heterogeneous size distribution, which the SSL had always shown and possibly determine the cause of this heterogeneity.

To make it more compatible for Tat-dependent selections, namely prevent the translocation of unfolded polypeptides, we devised the incorporation of a module encoding a patch of five consecutive hydrophobic amino acids.

Further, we chose to create an alternative to the SSL in the form of a completely unbiased, fully random library, which we deemed a valuable tool in the search for truly novel proteins.

2.2.1 Troubleshooting the huge size distribution of the SSL2.1

The SSL had been constructed in a randomly shuffled arrangement of individual modules with a limited alphabet of encoded amino acids and their respective codons, leading to a high sequence similarity on the DNA level for the variants of every single module.

Due to a completely modular construction scheme and a high sequence similarity within individual modules, sufficiently long sequence stretches of the DNA can easily hybridize with variants of one module type. This hybridization energy is high enough to cause undesired priming during PCR as well as robustly link multiple dsDNA products by inter-strand hybridization, leading to the very broad size distribution of the SSL observed on DNA agarose gels, which was also reported earlier [54].

The products obtained in the amplification of the SSL using different DNA polymerases, especially with increased 3'-5' exonuclease activity, suggested the following mechanism: single stranded DNA overhangs with a free 3'-end get efficiently excised when a module close to the 3'-end is able to hybridize with a similar module. All non-paired DNA on the 3'-end is quickly removed by the rapid exonuclease activity of the DNA polymerase. This inter-strand hybridization of similar variants of one module and the removal of unpaired 3'-ends by the DNA polymerase results in a shortening of the whole library with increasing PCR cycles, as short templates are generally amplified more efficiently.

The huge size distribution of SSL2.1 could be reduced greatly by purification of the sharpened band from an agarose gel under denaturing conditions (**Figure 2.17**) and PCR amplification using the Vent DNA polymerase. The PCR product of this amplification formed a relatively sharp band, even on non-denaturing agarose gels.

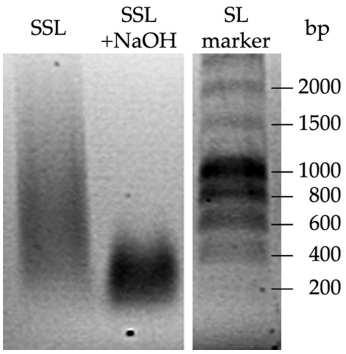


Figure 2.17: Ethidium bromide stained agarose gel of SSL2.1 under non-denaturing and denaturing (+NaOH) conditions.
The highly heterogeneous size distribution of the SSL2.1 that can be seen in the left lane, under non-denaturing conditions, is mostly due to hybridization of similar modules on different DNA stands. Mildly denaturing conditions caused by the addition of NaOH loading buffer eliminated most of these hybridizations between different members of the library and lead to a more compact size distribution, as seen in the middle lane;
size of SL marker bands indicated in base-pairs (bp).

2.2.2 Design of a hydrophobic patch library

After resolving the heterogeneous size distribution of the SSL2.1, the PCR product generated for cloning was found to be too short to encode for ~100 amino acids. Therefore, we combined two SSL2.1 entities and additionally incorporated a module encoding for a hydrophobic patch of 5 consecutive amino acids, to hinder the export of short unstructured polypeptides [106].

The PDBeMotif database (http://www.ebi.ac.uk/pdbe-site/pdbemotif/) was searched by pattern match on PDB sequences for patches of 5 consecutive amino acids of either Isoleucine, Leucine, or Valine (query: [ILV][ILV][ILV][ILV][ILV]). 1655 hits were found with the following total composition: 20% Ile, 40% Leu, and 40% Val. The DSSP (Define Secondary Structure of Proteins) [145,146] code classified 71% as part of extended β-sheets and 25% as part of α-helices. Initially, two consensus designs were generated, one with the highest propensity of encoding an alpha helix (LLVLL) and one for a beta strand (VVVVV).

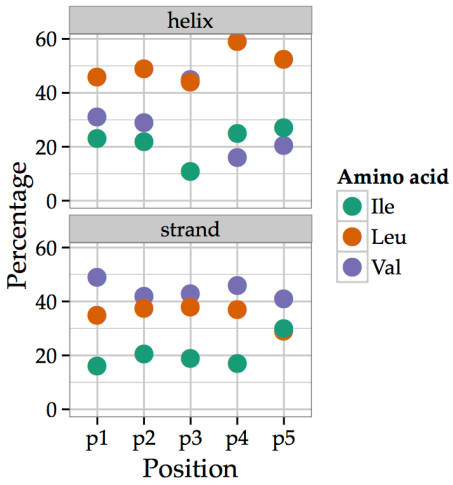


Figure 2.18: Secondary structure related analysis of patches of 5 consecutive hydrophobic amino acids found by PDBeMotif.
Distribution of the residues Ile, Leu, and Val according to the position in the stretch of five consecutive residues for sequences found in alpha helices and in extended strands.

The consensus designs were not pursued further in favor of a short library module Φ (phi), which encodes any of the three hydrophobic amino acids (Isoleucine, Leucine, and Valine [ILV]) with equal propensities in each of the 5 positions. The Φ module was realized with the degenerate/wobble codon VTH.

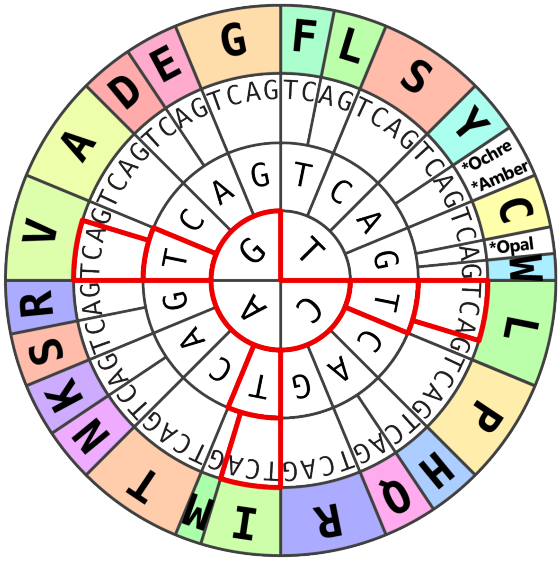


Figure 2.19: Design of the hydrophobic patch library Φ, encoding 5 consecutive hydrophobic amino acids by VTH wobble codons.
The three hydrophobic amino acids Ile, Leu, and Val are represented with equal propensities.

2.2.3 SSL-Φ-SSL cloning

To combine the two SSL2.1 entities with the central Φ (phi) module the following cloning scheme was employed: PCR1a with the forward primer fw_SSL_BclI on SSL2.1 and the established reverse primer cham5r yielded ΦSSL and PCR2 with the forward primer fw_NcoI_SSLjoin and the reverse primer cham5r yielded jSSL. The product of PCR1a contains at

the 5' end of the coding strand, the BamHI-compatible restriction site BclI, no NcoI site, the short library module Φ for the hydrophobic patch, and a BamHI site at the 3' end. Ligation of the BamHI digested PCR1a and BamHI digested SF·GFP-ssrA yielded ΦSSL-SF·GFP-ssrA, which was further amplified in PCR1b using fw_SSL_BclI and a reverse primer specific for SF·GFP-ssrA, rv_HindSacSsrA. The product of PCR1b was digested with the restriction enzyme BclI, and the PCR2 product jSSL with BamHI, generating compatible overhangs. Ligation of jSSL(BamHI) and (BclI)ΦSSL-SF·GFP-ssrA generated jSSL-Φ-SSL-SF·GFP-ssrA (**Figure 2.20**), which has neither a BclI, nor a BamHI recognition site before the Φ module. The ligation product was extracted from an agarose gel and further amplified using the outer primers fw_joinSSL & rv_HindSacSsrA.

The size of the library SSL-Φ-SSL was ~300 bp, encoding for 100 amino acids. This again showed the massive shortening of the SSL2.1 to less than 150 bp on average, as two SSL2.1 entities were combined for SSL-Φ-SSL.

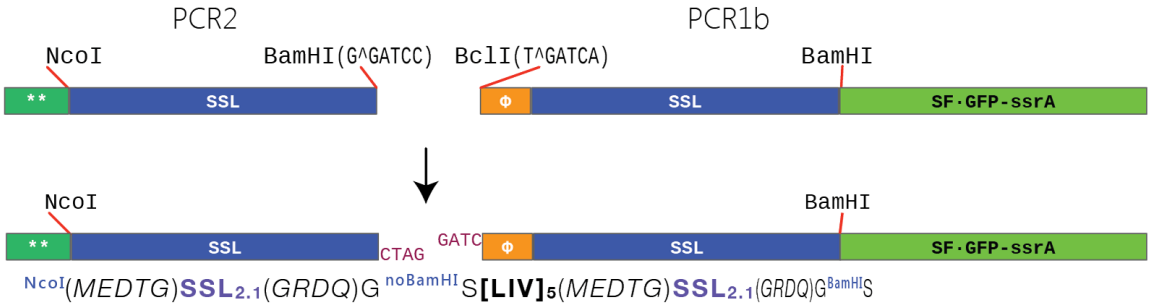


Figure 2.20: SSL-Φ-SSL construction.
PCR1b was digested with the restriction enzyme BclI, PCR2 with BamHI, generating compatible overhangs. The resulting ligation seam before the Φ module has the sequence GGATCA, which is neither a BclI, nor a BamHI recognition site.

2.2.4 SSL3 cloning: SSL-Φ-SSL with BamHI & PspOMI sites

SSL-Φ-SSL was re-cloned into the format of the POIs with flanking BamHI & PspOMI restriction sites to improve the reduced export rate, by avoiding the methionine after the signal sequence (see 2.1.12), encoded by the NcoI site that had been used for SSL2.1 cloning.

SSL3 was constructed by PCR amplification of SSL-Φ-SSL using the primers fw_SSL_noNco_TorA and rv_SpSPCRlib1.

In SSL3 the NcoI site was eliminated by changing its methionine codon to a threonine codon using the forward primer, which also introduced the 5' BamHI site. The 3' BamHI site was eliminated by the reverse primer, which kept the Glycine and Serine in the amino acid sequence by altered codons, and introduced the PspOMI site.

After amplification with the flanking primers SSL3 again showed a shortening of the library to ~250 bp.

2.2.5 Iterations of random library construction

The SSL only covers restricted regions in sequence space due to its binary patterning design. The high sequence similarity within each module causes many undesired side-products during PCR amplification and eventually leads to continuous shortening of the library.

As an alternative to the SSL, a completely random library was planed, which would make no assumptions regarding the secondary structure composition and no restrictions regarding the available amino acid alphabet. This random library should be seamlessly assembled and also incorporate the hydrophobic patch library Φ (see 2.2.2) to hinder export of unfolded polypeptides.

We decided to construct the random library from degenerate oligonucleotides using conventional cloning strategies like PCR, endonuclease restriction digest, and ligation. Although directed chemical coupling of oligonucleotides would reduce the experimental steps needed and possibly generate a greater yield and diversity, this approach has so far not been reported for successful assembly of libraries coding for ~100 amino acids.

The first design of the oligonucleotide encoding the main module for random library construction was rather complex. For the codon NNN the distribution of encoded amino acids does not change if the reading frame is shifted. In contrast, NNK codons would change to NKN in frame+1 or KNN in frame+2. NNN codons were used in the first design, as the ligation efficiency, especially the percentage of in-frame ligation products, was not known. To assemble a library encoding for ~100 amino acids, multiple building blocks have to be combined by restriction digest and ligation. If there is a considerable propensity of frame-shifts in each assembly step, a final product may still be in-frame but have one or multiple building blocks with frame-shifts.

To seamlessly assemble the random-library building blocks, the ligation overhangs have to be kept random as well. Otherwise the encoded amino acids at the ligation site would be fixed to one or a small set of possible residues. Type IIS [147] restriction endonucleases, which cut the DNA at a defined distance from their recognition sequence, can overcome this limitation. By positioning the recognition sequence of the type IIS restriction enzyme at the designated distance and orientation to the wobble codons, the cutting site can be placed in the randomized region of the building block.

2.2.6 Random library construction: version 1

For the first version of the random-library oligonucleotide, we searched for type IIS restriction enzymes with high specificity and a preferentially long distance between recognition sequence and cutting site, where the restriction sites for the selection vector could be incorporated. The file embossa_e.909 (September 2009), containing a list of all known restriction enzymes, was downloaded from the Restriction Enzyme dataBASE [148] (http://rebase.neb.com/) and searched for type IIS restriction enzymes having a recognition sequence of 6 or more base pairs and at least 8 bp distance to their cutting site. Features like stability, star-activity, over digestion, ligation and re-cutting efficiency were taken into account. BpmI and BseRI were chosen and their recognition sequences incorporated in the design of the first version of the random-library oligonucleotide.

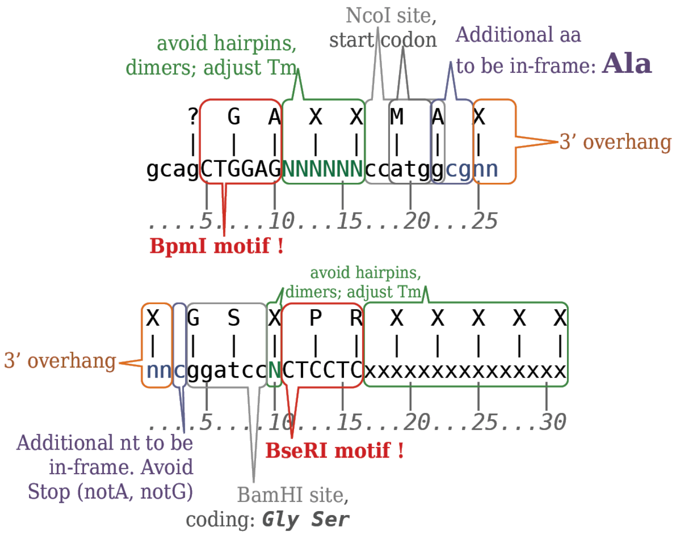


Figure 2.21: Design considerations for the regions flanking the randomized stretch in version 1 of the random library design.
Recognition sites for BpmI and BseRI are outlined in red, their cutting sites, the 3' overhangs are outlined in orange. The relatively long distances between recognition and cutting sites was used to incorporate various features, such as the NcoI and BamHI sites, which could later be used for cloning the insert into selection vectors.

The vector restriction sites (NcoI and BamHI) were placed in the sequence stretch between the recognition site and the cutting site of the type IIS endonucleases and thereby close to the random sequence. This has the advantage that the random library may be cloned into the selection vectors without having flanking peptide linkers of a few fixed residues. For the part with the BpmI site only the first nucleotide of the codon flanking the random sequence was fixed, having a Guanine nucleobase by the NcoI motive. The codon GNN could encode for any of the following amino acids: V, A, D, E, or G. It was set to GCG, encoding for Alanine. The 3' BamHI site is immediately after the random sequence codons. The remaining nucleotide positions were used to optimize melting temperature and G/C content besides minimizing hairpin and dimer formation.

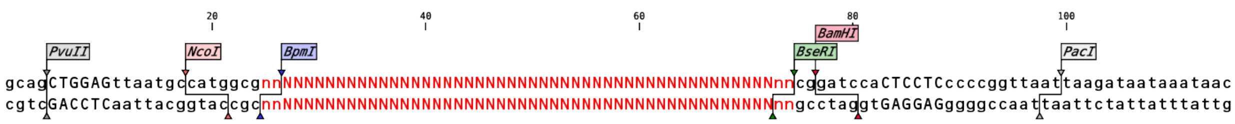


Figure 2.22: Double stranded sequence of random-library oligonucleotide version 1 after fill-in PCR.

For the oligonucleotide encoding the hydrophobic patch library the choice of the type IIS restrictions enzymes had to be adjusted, as a BseRI recognition sequence would be present with a propensity of 4/81 (~5%) in the five VTH codons, leading to a shortening or loss of the hydrophobic patch. The BsrDI restriction site, generating compatible overhangs for ligation, was used instead.

All construction steps of the random library were checked in silico to ensure the reading frame was retained, particularly for the hydrophobic patch module. The fill-in PCR reaction and the amplification of the obtained double-stranded DNA with outer primers worked quantitatively. The type IIS restriction digests cut more than 90% of the template, when the recognition site was

flanked by at least 30 bp on both sides. Purification of the cut fragments was tested using PAGE and agarose gel electrophoresis. Agarose gel extraction combined with column purification was found to deliver the highest DNA recovery.

However, the ligation reactions using two cut and purified fragments, one digested with BpmI and the other with BseRI, yielded only very low amounts of the desired ligation product (less than 5% of input DNA). PCR amplification with outer primers on the gel-extracted band resulted in multiple products, the largest one of the size of the desired product. Running the DNA of this band under denaturing conditions on an agarose gel resulted in its separation into two smaller fragments and the disappearance of the desired band, suggesting that the fragments were not covalently linked.

The overhangs after digestion were chosen to be random, theoretically allowing the non-directed ligation of two identical fragments or of the two different fragments. Yet, no products corresponding to the ligation of two identical fragments were observed.

The purified product of the ligation reaction may have been merely a hybridization product rather than the covalent fusion of both fragments. The ligation efficiency was very low, conceivably because the chosen restriction enzymes generate an NN-3' overhang of only two nucleotides. Trials where these digested fragments were additionally blunted prior to ligation resulted in even lower yields of the desired (putative) product. The chosen type IIS endonucleases were uncommon and commercially available only in low concentrations, further complicating the construction of a full length library at large scale.

We therefore revised our strategy and addressed many of these issues in the second version of the random library design.

2.2.7 Random library construction: version 2

To overcome the inefficient assembly of individual library modules, a more reliable type IIS restriction enzyme, BsaI, was chosen for the second version of the random-library oligonucleotide. BsaI sites were located at both ends of the random sequence stretch. Using a pair of primers, which introduce a point mutation in the recognition sequence, each BsaI site can be deactivated or (re)-established. In this design, there was only one extra nucleotide between the recognition sequence and the cutting site of BsaI. The restriction sites for cloning the library into a plasmid vector were therefore moved to the outer regions of the BsaI sites and changed to the format with higher export rates (BamHI and PspOMI, see 2.1.12). The codons for the random sequence were kept as NNN.

Figure 2.23: Double stranded sequence of random-library oligonucleotide version 2 after fill-in PCR.

The oligonucleotide for the random library was converted to double stranded DNA in a fill-in PCR and further amplified with flanking primers. This PCR product was subsequently fused to

DNA modules of different lengths to allow a better separation of the different ligation products expected from the non-directed ligations (see 2.3).

With the four nucleotide overhang generated by BsaI in the randomized sequence, the modules used for ligation all contained a 5'-NNNN overhang.

Due to the randomized overhangs, ligations of the fragments are not directed. Each 5'-NNNN overhang could hybridize with a complementary 5'-NNNN overhang from an identical or different fragment.

Two variations of the random module were generated by PCR, one with an active BsaI site at the 3' end, the other with an active BsaI site at the 5' end. We observed that the 5' BsaI site, which is present in the long random-module oligonucleotide, could not be inactivated by the initially chosen primer. This primer was designed to introduce a point mutation in the BsaI recognition sequence and thereby inactivate it. The mutation is located in the second last nucleotide of the 3' end of the primer. All constructs generated with this primer still carried an intact BsaI site. Most likely the high proofreading and 3'→5' exonuclease activity of the employed DNA polymerase lead to the removal of the last two nucleotides of the primer, including the mismatch designed to inactivate the BsaI site. A revised primer for the inactivation of the BsaI site, carrying a mismatch nucleotide further away from its 3' end, worked successfully and mutated the BsaI recognition sequence.

Ligation of the cut fragments, fused to DNA modules of different lengths, showed the expected products of a non-directed ligation due to the 5'-NNNN overhangs. This time, all the possible intra- and inter-module ligation products were observed. The ligation efficiency was still low at about 5-10% of input DNA. The desired inter-module ligation product was gel extracted, amplified by outer primers, and cloned into a plasmid vector. Sequencing of the PCR product and a few single clones confirmed the successful seamless joining of two random modules. Sequencing of the single random-library module and the two modules joined by ligation both showed the designed number of nucleotides (**Figure 2.34**), indicating that the used oligonucleotide was of good quality and that the ligation of overhangs composed of four random nucleotides worked thoroughly, regarding the very low occurrence of frame-shifts. This was supported by analysis of single clones.

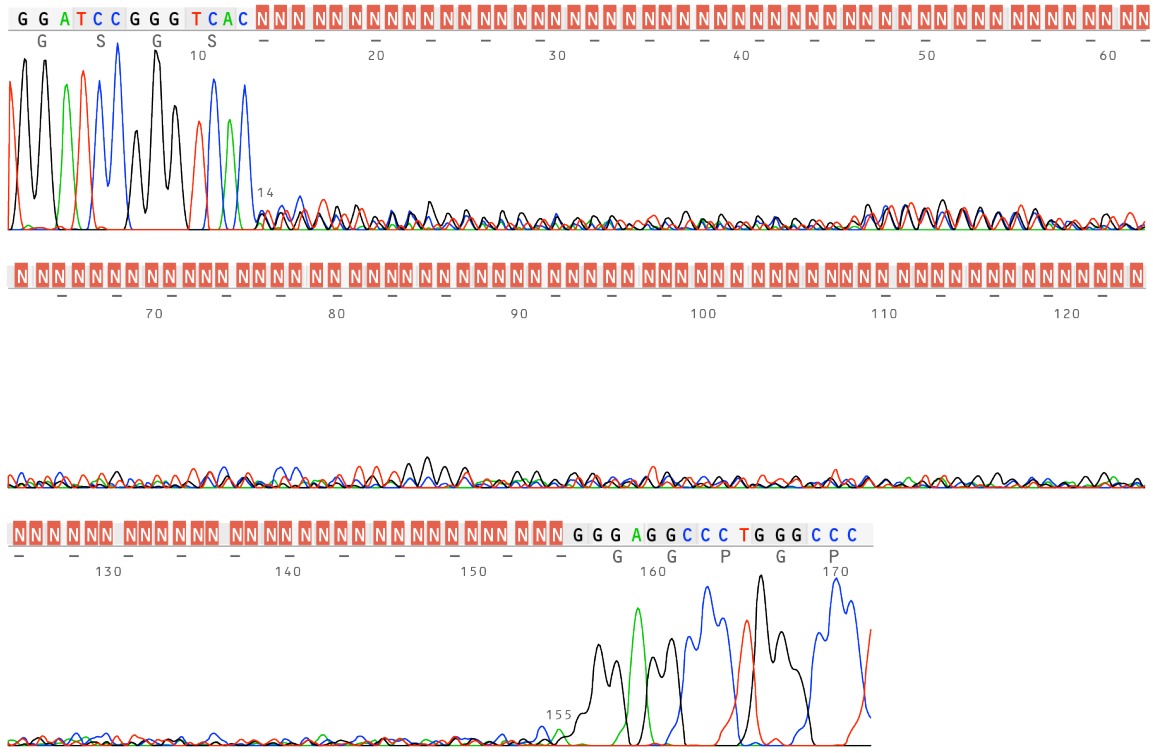


Figure 2.24: Sequencing chromatogram of the gel-extracted ligation product of two modules. The traces show the defined sequence stretches flanking the randomized region, beginning with the BamHI site (GGATCC) and ending with the PspOMI site (GGGCCC). The randomized region consists of the designed number of 142 "N" nucleotides, with the first codon being CNN and the last NNG, encoding 48 randomized position.

2.2.8 Random library construction: version 3

For the third version of the random-library oligonucleotide the codons for the random sequence were changed to NNK. Ligation products of two random modules of the second version showed almost no frame-shifts and the main reason of using NNN codons initially was the concern of a higher propensity of insertions or deletions due to incorrect hybridization and ligation of restriction digest overhangs. The sequence surrounding the randomized codons was kept similar to the second version, including the flanking BsaI sites on both ends. As before, one single oligonucleotide containing the random library module was used for assembly.

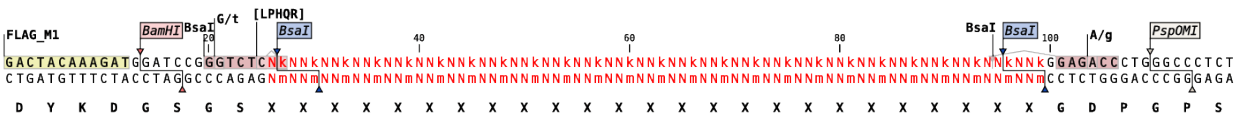


Figure 2.25: Double stranded sequence of random-library oligonucleotide version 3 after fill-in PCR.

The steps up to the construction of two joined random modules worked comparable to or better than previous versions. Fixed sequence DNA modules were fused to the random module to make a separation of the different products of the non-directed ligation feasible. Using outer primers for the fixed DNA modules, the whole ligation product, containing the two joined random modules, could be amplified by PCR.

Yet, whenever a primer was used for re-introducing a BsaI site, flanking the randomized codons, the main PCR product contained only one random module. PCR products containing two joined random modules were barely visible or not detectable on gel, when using a primer close to the randomized sequence. The hybridization sequence of these primers is located inside the single random oligonucleotide and thus is present in a single module as well as two joined modules. Tiny amounts of co-purified template containing only a single module will therefore be preferentially amplified, if primers are used that can bind within this sequence.

The successful assembly of a full length random library was achieved using the forth and final design. There, two distinct oligonucleotides were employed that included, in addition to BsaI, recognition sites for a second, reliable, type IIS restriction endonuclease, BpmI, generating overhangs compatible to the BsaI digests.

Furthermore, the two oligonucleotides contained signature sequences flanking the randomized nucleotides. These signature sequences comprised the type IIs recognition sites and the adjoining fixed sequences. They were designed to have very low similarity to each other and thereby enable the specific and exclusive amplification of ligation products consisting of joined modules even when using primers close to the randomized sequence.

A detailed description of the complete construction procedure for the full length random library can be found in the next section.

2.3 Successful construction of the MOAL, a fully random library of 303 bp

The following sub-chapter explains the strategy and construction of a seamlessly assembled, completely random, full length library encoding random proteins of 101 amino acids, which we termed MOAL. This sub-chapter has been written in the form of a small paper draft with the intention that it could as well be read on its own, mostly independent of the other (sub-)chapters of this work. Therefore, it accompanies its own introduction, results, and discussion parts, which may partially coincide with certain sections of the general chapters.

2.3.1 Abstract (MOAL)

The search for truly novel proteins requires libraries of modules encompassing sequence space without homology to natural proteins or of completely random composition. By design, such a library can additionally incorporate features that may serve as nucleation center for folded proteins.

Here, we report the successful construction of a fully random DNA library with a final length of 303 base pairs using oligonucleotides and Type IIs restriction enzymes. This DNA library is composed solely of NNK codons, except for a central patch of 5 VTH codons (coding for Val, Leu, and Ile), which was included to encode for a hydrophobic patch in the translated proteins, and can serve as nucleation core for folded proteins.

To seamlessly assemble the building modules of the library the overhangs of the Type IIs restriction enzymes were kept random, not permitting a directed ligation as each module could hybridize with any other module.

The construction modules were therefore fused to DNA-stretches of distinctive lengths making the different ligation products separable by size. Signature sequences allowed the specific amplification of the desired ligation products.

By the use of high-fidelity PCR, robust Type IIs restriction enzymes, and T4 DNA ligase a high quality of the final library was ensured, without the need of pre-selections on the individual modules.

Sanger sequencing of more than 300 single clones verified the solid quality of the final library with regard to being in-frame, codon composition and concordance to design (e.g. containing the central patch of 5 VTH codons).

The whole library was assembled completely on DNA level. It is ideal for exploring sequence space, which is not covered by previous, biased libraries or libraries with translational selection steps in their construction.

2.3.2 Introduction (MOAL)

How can one scout unexplored sequence space for truly novel proteins?

For an experimental, non-computational approach, the challenge is to build a useful library of the size of a typical protein domain encoding about 100 amino acids [46-48]. This library should mainly cover unexplored sequence space and be of high quality, containing a very low number of frame-shifts and stop codons. If there is no way to predict which limited amino acid composition would have a higher propensity of obtaining folded structures, this library should further be as unbiased as possible.

Only few examples of random libraries coding for polypeptides longer than 40 amino acids have been reported [149-152]. Experimentally, it is already not feasible to fully sample a library of an 11 amino acid polypeptide allowing all 20 standard residues in each position [49,150], which has a theoretical diversity of $20^{11} \approx 2 \times 10^{14}$.

Unrestricted libraries encoding longer polypeptides will surpass the limit that can be screened experimentally by far and it won't be possible to encode and select a unique sequence in such a huge potential diversity. The goal must thus be to encode an ensemble of sequences sufficiently similar to useful sequences that fold stably, such that the truly functional ones can then be obtained under evolutionary pressure, e.g. by scouting the local sequence space with introduction of further point mutations.

Regarding the sampling limitations, a decision has to be made in the design of a library to either stay within or close to the explorable limit of about 10^{14} different elements, that can still be fully sampled (regarding the translation to proteins and passing through a selection system) or to give up this requirement for a higher theoretical diversity. In the first case, it is crucial to make sure that the encoded diversity actually contains the solution. In other words, if truly novel proteins shall be discovered, it has to be certain that they can be retrieved from this limited sequence space. In the second case, when using a library with a theoretical diversity *above* the explorable limit (i.e. $\sim 10^{14}$ different elements) only a small subset can be sampled at one time. One

underlying idea for such an approach is that it might be sufficient to land close to an interesting region in sequence space to stand out of the bulk – and from that point on further improvements can be made to reach the optimal point in sequence space (e.g. stable folding) – which, by that notion, does not have to be in the initially sampled subset of the library [13,153].

The great advantage of a library that by design can be fully sampled experimentally is that the contribution of each individual constituent can be mapped (e.g. mapping the distribution of selected amino acid residues to the applied selection pressure). Moreover, an optimum could be approached that may not be present in the library itself by combining characteristics of different favorable, mapped elements [154].

Limiting the theoretical diversity, so that a library can still be sampled experimentally, might be achieved by using only a subset of different amino acids [152], by employing certain patterns of characteristic amino acids or groups of amino acids, and by combining various (established) modules.

Restricting the number of different residues along with attributing certain properties [155,156] lowers the theoretical complexity of the system and allows longer polypeptides, while sustaining an explorable diversity – yet, even when limited to just five different residue types, a polypeptide of 21 amino acids would already have too many potential combinations ($>4\times10^{14}$) to get fully sampled.

Moreover, such restrictions severely trim down the sampled sequence space in a way that only small clusters are populated, which may be too distant to the spots one is interested in (e.g. being well-folded proteins).

Another approach to limiting the available residues for the whole stretch is to define distinct patterns, where the allowed residues are restrained for each position individually [51]. While single modules may still be fully sampled in their total diversity, by combining multiple modules (with different patterns) the theoretical diversity increases rapidly beyond the explorable limit.

Combination of binary patterned modules [51] have been proposed to increase the probability of forming e.g. secondary structures [157].

As the theoretical sequence space can be immensely large compared to the number of variants that can be handled experimentally, one has to carefully consider the implications of limiting the diversity of library design. In most cases one very important consideration is the probability of encoding well-folded proteins; as it is typically the protein function one is interested in, which is closely coupled to its folding.

One tempting approach to increase the likelihood of encoding well-folded proteins is to use sequence stretches [158,159] or sequence compositions homologous to naturally occurring proteins [160-162]. However, this largely restricts the covered sequence space and most likely leaves out most of the sequence space that possibly has not been sampled by nature/evolution.

Biased (random) libraries have also been constructed to resemble the composition of natural proteins, mostly covering a large theoretical sequence space beyond the explorable limit [150,163].

The most prominent example of an almost fully random library is the one by Cho et al. coding for 80 amino acids, which was successfully used for selections of well-folded proteins binding to ATP [44].

It is composed of a pre-selected module (coding for 20 amino acids) fused four times by directed ligation, leaving the encoded amino acids at the ligation seams not fully random, but with few possible residues. Notably, in the course of selections two residues in the ligation seams were changed to amino acids not encoded in the design, possibly hinting that the restriction of encoded amino acids in the ligation overhangs may have been disadvantageous.

Some features of natural proteins, such as the solvent-shielded hydrophobic core [153], are challenging to recreate in libraries using translational steps in their construction. To scout unexplored sequence space for truly novel proteins, modules that serve a central function in a full-length protein, but have unfavorable properties as short isolated module, should not be left out (e.g. a stretch of hydrophobic residues forming the core of many proteins).

Approaches to constructing a random library include the use of trinucleotide codons [53,125,164] and standard oligonucleotides with mixed bases at all or certain positions.

Trinucleotide codons allow a well defined codon mixture for each position [165] and frame-shifts (to frame+1 or frame+2) occur potentially less frequently since the DNA is assembled codon by codon. However, the coupling efficiency is lower than for mononucleotides, and no trinucleotide libraries coding for more than 20 consecutive amino acids have been reported [53].

The use of degenerate oligonucleotides with wobble bases encoding the mixture of amino acids has the advantage that it is a well established method generating high yield of a good quality library regarding the control of the final length and composition of bases. The wobble codon NNK is a good compromise for encoding all amino acids with reasonable weights (only the propensity for I, M, and W change compared to NNN) [166] while losing two stop-codons (keeping only the amber-stop TAG which could be read-through e.g. in a *glnX* a.k.a. *supE* background) see **Figure 2.26**.

The quality of the library depends on the quality of the oligonucleotides and the assembly procedures of the library. In the worst case, the fraction of in-frame sequences can be as low as 33%, leading to 67% change of encoded amino acid composition due to frame-shifts (except for codons with the same mixture of bases at each position, e.g. NNN, where frame+1 and frame+2 still consist of NNN codons). Often translational steps ("in-frame selection") have been used on each module to enrich in-frame sequences [136,137]; yet, this again reduces the diversity and eliminates sequences that have unfavorable characteristics as isolated modules but might be useful as a component of the final library.

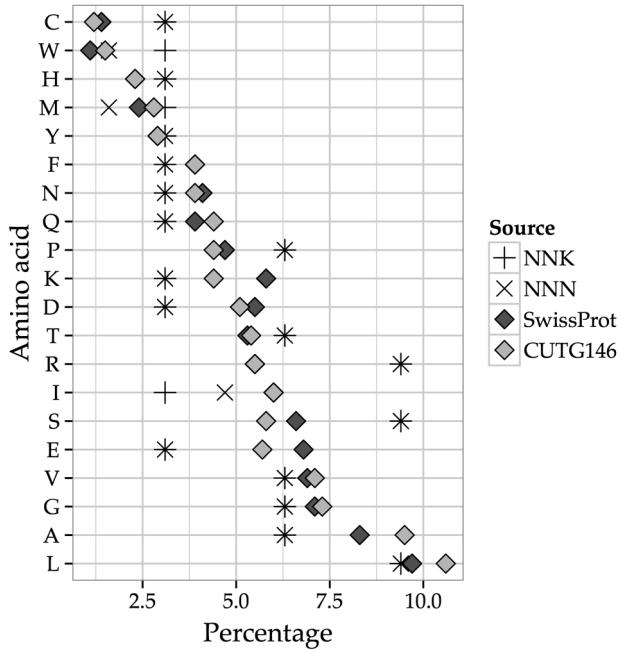


Figure 2.26: Expected amino acid distributions for NNN or NNK codons compared to natural proteins. SwissProt: Amino acid composition (%) in the UniProtKB/Swiss-Prot data bank as found in the release notes for UniProtKB/Swiss-Prot release 2013_04 - April 2013 [167]. CUTG146: Translated codons of 5045 coding sequences of *Escherichia coli* K12 (Division: gbbct, Release: CUTG146) as found in e.g. EMBOSS [168].

Here we chose to keep the design as unbiased as possible and probe unexplored sequence space in the best way possible by using NNK codons. We included a feature that could hardly be incorporated when using any intermediate translational steps, namely a central patch encoding for five consecutive amino acids. The joints of the module were kept random as well (with KNNK ligation overhangs) in order not to limit any position to encode for unfavorable (sets of) amino acids, which would be common to all library members. We describe here the design and assembly strategy of such a library encoding 101 amino acid positions.

2.3.3 Materials and Methods (MOAL)

Oligonucleotides were ordered at Microsynth, the ones longer than 40 nucleotides in PAGE purified form.

Table 2.3: Oligonucleotides used for amplification of the respective library modules.

BpiNNKBb	CTTATTTCTTCTGTGTGCGAAGACGNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNK NNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKGGAGGCCCTGGGCCCTCT
rv_NNKb	GTTCTTCTCCTTTGCTACCAGAGGGCCCAGGGCCTCC
bNNKBpi	GACTACAAAGATGGATCCGGTTCTCNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNK NNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKCCGTCTTCTCGCCTGCCG
rv_NNKBpi	CACTAAATGTTTACGCTAACACGGCAGGCGAGAAGACGG
phiNK	CCGGTCTCNKNNKVTHVTHVTHVTHVTHNNKNNKCGGTCTTCTAACTGAGTAACGTCAGG TGACCGGG

Oligonucleotide names are given in the left column, their 5' to 3' sequences in the right column.

PCRs were performed using the Phusion DNA polymerase (Finnzymes/NEB) and ligations using the T4 DNA ligase (Fermentas), all steps according to the suppliers' manuals.

The restriction enzyme BsaI was ordered at New England Biolabs and BpiI at Fermentas International Inc.

DNA agarose gels were prepared using 0.5×TBE and run at a constant voltage of 110 Volts (10 V/cm).

PCR purifications and gel extractions were carried out using the designated kits from Machery-Nagel and/or QIAgen. Purifications of restrictions enzyme digests were also done using the PCR purification kits. Column purification and gel extraction of the same amount of input DNA were compared regarding the final quantity of desired ligation product and the feasibility of its separation from undesired side-products on an agarose gel; column purification of the restriction digests was found to yield significantly higher amounts of the desired ligation product. Digests were checked on agarose gel for completeness.

The final library was cloned via BamHI & PspOMI (flanking the random library) into a sequencing vector.

Sanger sequencing of single clones was performed by GATC Biotech using the G/C rich option for the sequencing reactions, as most reaction would otherwise give reads shorter than ~300 bp, not spanning the whole length of the library stretch.

2.3.4 Results (MOAL)

The goal was to create a random library encoding for about one hundred amino acids. This library should include a patch of five consecutive hydrophobic amino acids and, if possible, get seamlessly assembled. We decided to use NNK codons for the fully random parts and VTH codons for the hydrophobic patch, which thereby will be composed of Leu, Ile, and Val with equal propensities. One major challenge was the seamless assembly using ligation overhangs located in the randomized region. This non-directed ligation resulted in products composed of any combination of educts. To make it feasible to obtain only the desired ligation product, the different possible ligation products were designed to differ substantially in size and signature sequences were included, which allowed the exclusive amplification of the desired ligation product after gel-extraction.

Generation of initial NNK library modules:

The use of NNK codons approximately retains the distribution of encoded amino acids as a fully random NNN design would have, but instead of 3 stop-codons NNK only has one (TAG), which can also be suppressed in e.g. *glnX* a.k.a. *supE* strains of *E. coli*. Thus, the probability of a stop codon drops from 3/64 to 2/64 (=1/32) and 4.7% of the encoded polypeptides of 101 amino acids (96 NNK + 5 VTH codons) will have no stop codons. Nonetheless, a frame-shift would significantly alter the codon composition, and therefore a very high oligonucleotide quality and precise assembly are fundamental.

Oligonucleotides BpiNNKb and bNNKBpi (both 117 nt) were designed to serve as building blocks for the NNK modules, each containing 24 NNK codons, a flanking BpiI site (BpiI: GAAGACNN^nnnn) on either the 5' or 3' end, and a defunct BsaI recognition sequence (BsaI: GGTCTCN^nnnn) with one mutation on either the 3' or 5' end (denoted as "b"), which can be reactivated later by reverting this point mutation.

Large-scale PCR amplifications of BpiNNKb with rv_NNKb as reverse primer, and bNNKBpi with rv_BpiNNK were performed to generate double-stranded DNA from the initial oligonucleotides encoding the NNK library (**Figure 2.27a**). This created the two random modules, each containing 25 randomized codons, which were then combined to form longer, seamless stretches of NNK codons.

The PCRs on the initial oligonucleotides worked in a quantitative way with respect to the starting template, i.e. the oligonucleotides containing the NNK codons. More than 95% of the long oligonucleotides coding for the random (NNK) library used in the initial PCRs were converted to double-stranded DNA, confirmed by a comparison of the oligonucleotide and the purified PCR product on an agarose gel.

Prerequisites for assembly of building blocks by non-directed ligation:

To generate a random library encoding for 101 amino acids, we combined multiple building blocks. In a first step, the two different NNK modules (25 randomized codons each) were fused to generate one long module (2N) containing 47 NNK codons. This "2N" module was later used again to complete the library after the incorporation of the phi module for the central hydrophobic patch.

For the ligation of two building blocks, we always used type IIs restriction enzymes with recognition sites flanking the randomized regions and cleavage overhangs that are located within the region of the NNK codons. These overhangs are hence composed of random nucleotides and allow to keep the ligation seams random. Consequently, a directed ligation is not possible, as each module can ligate with itself or its designated partner – the ligation overhangs were designed to be KNNK and would preferentially hybridize with MNNM on the complementary strand, but K (=T/G) can also base-pair with K (=G/T) as well as M (=A/C) can hybridize with M (=C/A), leading to non-directed ligation of all fragments: KNNK hybridizes with MNNM, but also KNNK with KNNK, and MNNM with MNNM.

To make a good separation of possible module combinations feasible, and to separate the wanted from the unwanted ligation products, the NNK modules were therefore fused to DNA modules of different lengths (tangerine, green and brown in **Figure 2.27**).

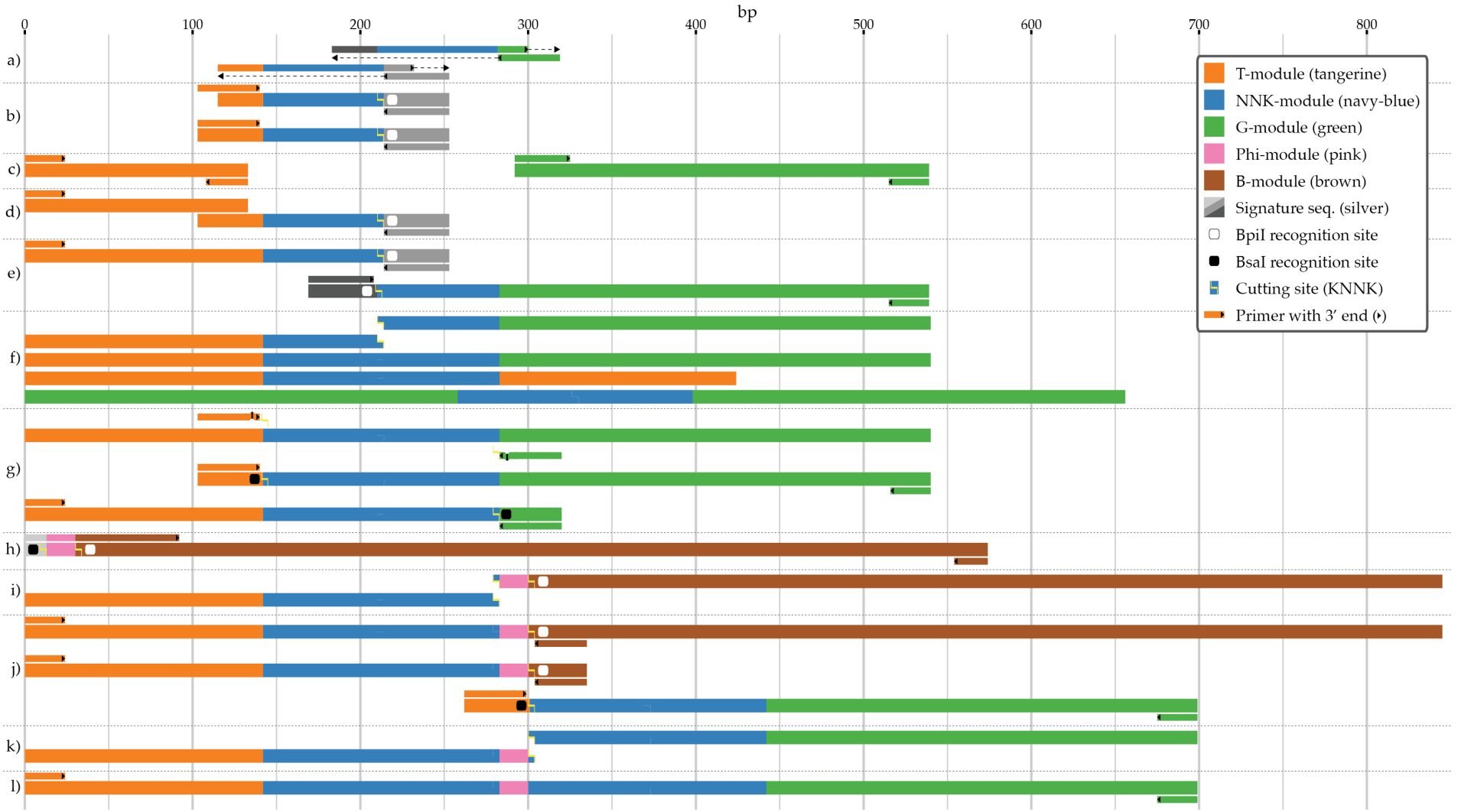


Figure 2.27: Cloning scheme of the library: a) fill-in PCRs of BpiNNKb with rv_NNKb and bNNKBpi with rv_BpiNNK; b) further PCRs for the generation of the NNK-module t-1N-BpiI; c) PCRs for the T-module and the G-module; d) PCR for fusing the T-module to t-1N-Bpi (NNK-module); e) PCR products Tmod-1N-BpiI and BpiI-1N-Gmod for BpiI digest and ligation; f) Ligation of BpiI digested fragments yielding the desired product Tmod-2N-Gmod and the two major side-products; g) primers flanking the 2N region introducing BsaI sites and the respective PCRs with outer primers; h) PCR for the central hydrophobic patch "phi" yielding BsaI-phi-BpiI-Bmod; i) Ligation of fragments after BsaI digest; j) Ligation product with primer flanking the hydrophobic library module and the two PCR products used for the last digest and ligation step Tmod-2N-phi-BpiI and BsaI-2N-Gmod; k) ligation of final fragments; l) final library with outer primers after gel extraction of the 699 bp ligation product.

To meticulously separate the different ligation products and facilitate gel extraction of the desired product, the B-, G-, and T-modules were designed to obtain fusion blocks of noticeably different sizes. Their exact sequence is only of minor importance as long as they don't share homologous sequence stretches or contain repeats of high similarity. Sequences of plasmids available in our laboratory were used here: B-module: part of a β-lactamase gene (540 bp), G-module: fragment of a GFP gene (256 bp), T-module: stretch of TorA signal sequence (142 bp).

Using adapter primers that introduce identical sequence stretches (see overlapping tangerine parts in **Figure 2.27d**), the T- and G-modules were fused to the NNK modules to generate modules of distinct lengths for the non-directed ligation. **Figure 2.27b** shows the construction of one of the NNK modules (t-1N-BpiI) from the PCR product of bNNKBpi and rv_BpiNNK (lower part of **Figure 2.27a**), which was amplified using fw_NNK_t & rv_NNK_Bpi to yield t-1N-BpiI. The T-module (Tmod) was then fused to this NNK module (t-1N-BpiI). The construction of the other NNK module was performed in the same way, where the PCR product of BpiNNKb and rv_NNKb was amplified using fw_BpiNNK & rv_NNKb yielding BpiI-1N-g, the NNK module to be fused to the G-module (Gmod). For this NNK module (BpiI-1N-g), the individual construction steps are not shown in **Figure 2.27**, yet the lower part of **Figure 2.27e** shows this NNK module fused to the G-module, BpiI-1N-Gmod.

Figure 2.27c shows in tangerine color the 133 bp T-module and in green color the 248 bp G-module, which were generated in separate PCRs with outer primers. For T-mod, fw_Tmod & rv_Tmod were used, and G-mod was amplified with fw_Gmod & rv_Gmod.

The PCRs for fusing the random (NNK) modules to the designated "constant" DNA modules (T, G, or B) were performed by overlap extension PCR, i.e. by mixing the respective purified PCR products in equimolar ratio and adding outer primers.

Figure 2.27d shows the overlapping region of 30 bp of Tmod and t-1N-BpiI in tangerine color, and further the outer primers fw_Tmod (above, in tangerine) & rv_NNK_Bpi (below, in silver), which were used to amplify the fusion product Tmod-1N-BpiI of 253 bp (see upper part of **Figure 2.27e**).

Similarly, the other NNK module (BpiI-1N-g) was fused to the G-module to obtain BpiI-1N-Gmod of 370 bp (see lower part of **Figure 2.27e**). The individual steps are not shown in **Figure 2.27**. BpiI-1N-g and Gmod, having a 28 bp overlap, were mixed with addition of the outer primers fw_BpiNNK & rv_Gmod resulting in the fusion product BpiI-1N-Gmod of 370 bp.

Both random modules, fused to their designated constant modules, Tmod-1N-Bpi and BpiI-1N-Gmod were further amplified using outer primers (**Figure 2.27e**).

These assembly-PCRs again worked in a quantitative way as assessed by the disappearance of the educt bands on agarose gels of the purified assembly-PCRs.

Non-directed ligation of NNK modules:

The upper part of **Figure 2.27f** shows the fragments created by BpiI digests of the respective library modules (**Figure 2.27e**). Theses fragments were used for the ligation of two NNK

modules: Tmod-1N-(BpiI) of 210 bp + 4 nt overhang (plus a smaller fragment of 39 bp + 4 nt) and (BpiI)-1N-Gmod of 326 bp + 4 nt overhang (plus a smaller fragment of 40 bp + 4 nt).

The BpiI restriction digest fragments were used in equimolar amounts as input for the ligation of Tmod-1N-(BpiI) and (BpiI)-1N-Gmod (upper part of **Figure 2.27f**) yielding the desired ligation product Tmod-2N-Gmod of 540 bp and multiple side-products; e.g. the educts as well as Tmod-1N-(BpiI) ligated with an identical module at 424 bp and (BpiI)-1N-Gmod with an identical module at 656 bp (**Figure 2.27f, Figure 6.1**). Only the desired ligation product Tmod-2N-Gmod of 540 bp, shown as the first of three ligation product in **Figure 2.27f** and composed of the tangerine+navy-blue+green modules, was gel-extracted and then amplified with outer primers fw_Tmod and rv_Gmod.

The step determining the maximal experimental diversity (i.e. amount of DNA coding for the random library) is the gel extraction of the ligated modules – in all other steps the retainable diversity is at least one order of magnitude higher. Estimation by spectrophotometry and by agarose gel ethidium-bromide staining of recovered DNA for the ligated modules allowed an approximate determination of the maximal diversity of the final library to be around 10^{13}.

To further use the two fused NNK modules (2N) for the incorporation of the phi module and the final assembly of the library the defunct BsaI recognition sequences were converted by flanking primers to introduce correct, functional BsaI sites, generating again KNNK overhangs. **Figure 2.27g** shows the location of the primers introducing the functional BsaI sites by a point mutation. Using fw_Tmod & rv_NNK_BsaI the 322 bp PCR product Tmod-2N-BsaI was generated. The other PCR product was obtained using fw_NNK_BsaI & rv_Gmod yielding 436 bp BsaI-2N-Gmod (**Figure 2.27g**).

Both flanking primers carry one mismatch each, which introduces a functional BsaI recognition site. For the construction of a "2N" module, as described here, one of the BsaI recognition sites could already be functional in the initial oligonucleotides. The use of defunct BsaI recognition sequences in the initial oligonucleotides, however, allows a higher flexibility in the construction, especially if more than two random modules were to be combined.

Incorporation of the phi module for the central hydrophobic patch:

Well-folded proteins often contain a solvent protected core composed of hydrophobic amino acids. But stretches of hydrophobic amino acids are hardly recovered as isolated modules using translational steps. As they can serve as nucleation region for forming the protein core, we included this feature in the center of our library. The key consideration in introducing this hydrophobic stretch was to prevent selection of natively unfolded, very soluble proteins.

The Φ (phi) module for the central hydrophobic patch was provided directly by the phiNK oligonucleotide, which contains five VTH codons, encoding a stretch of 5 hydrophobic amino acids, each with 3 codons: Leu (CTH), Ile (ATH), and Val (GTH).

The generation of the brown colored, 534 bp B-module (Bmod) by PCR using the flanking primers fw_Bmod and rv_Bmod is not explicitly shown in **Figure 2.27**. However, **Figure 2.27h**

shows how the phi module, was fused to the B-module by using the phiNK oligonucleotide, which contains the five VTH codons. The 68 nt oligonucleotide phiNK, shown in silver-pink-brown, was used as forward primer to directly join the phi module coding for the hydrophobic patch with the B-module, and in combination with the reverse primer rv_Bmod yielded the PCR product BsaI-phi-BpiI-Bmod of 575 bp (**Figure 2.27h**).

Tmod-2N-BsaI of 322 bp and BsaI-phi-BpiI-Bmod of 575 bp were both digested with BsaI, ligated (**Figure 2.27i**), and the desired ligation product of 845 bp was gel-extracted (Tmod-2N-phi-Bmod, shown as the upper most construct in **Figure 2.27j** in the colors tangerine+navy-blue+pink+brown).

Assembly of the full-length library:

Up to this point, the library was assembled in the following steps. First, two random modules were joined to generate the longer "2N" stretch, composed of NNK codons. Then the phi module, encoding a stretch of five consecutive hydrophobic amino acids, was seamlessly ligated to the "2N" stretch. From the perspective of the encoded proteins, this generated a N-terminal random stretch of 48 amino acids, followed by five hydrophobic residues, where each of the five positions contains either a Leu, Ile, or Val with a probability of 1/3.

To obtain the full length library, the modules encoding the C-terminal random stretch have to be added in a final assembly step. Therefore, we reused the "2N" module from the first ligation of the two NNK modules. The middle and lower parts of **Figure 2.27j** show how the final assembly blocks were obtained. Tmod-2N-Φ-BpiI of 335 bp was generated using the primers fw_Tmod and rv_phiNK. Tmod-2N-Φ-BpiI was fused to BsaI-2N-Gmod, thereby reusing the "2N" module from the first ligation of the two NNK modules (see **Figure 2.27g** for creation of BsaI-2N-Gmod).

Tmod-2N-Φ-BpiI was digested with BpiI and BsaI-2N-Gmod was digested with BsaI. Then both were ligated (**Figure 2.27k**) and the final desired ligation product Tmod-2N-Φ-2N-Gmod of 699 bp was gel-extracted and amplified with the outer primers fw_Tmod and rv_Gmod yielding Tmod-2N-Φ-2N-Gmod (**Figure 2.27l**). This contains the final full-length library with long flanking regions. These flanking regions allow an efficient amplification of the full-length library, similar to the signature sequences used earlier. They also contain the restriction sites needed to clone the library into established selection vectors.

Quality of the final library:

To assess the quality of the final library, we analyzed 328 members by Sanger sequencing: 203 clones (62%) were in-frame, 83 clones (25%) were frame+1, 38 clones (12%) frame+2, and 4 clones (1%) with an unrelated sequence.

Of the 203 in-frame sequences, 14 (7%) were without any stop codon, 165 (81%) of designed length (coding for 101 amino acids) and 161 of these 165 had the hydrophobic patch phi as designed, the other 4 had point mutations leading to a change of one of the encoded hydrophobic amino acids.

Figure 2.28 shows the distribution of insert lengths for the 203 (62%) in-frame sequences.

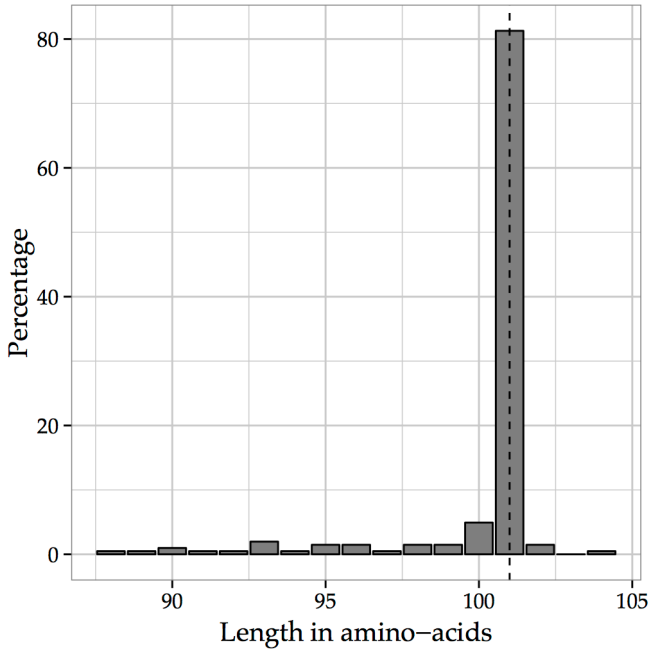


Figure 2.28: Distribution of insert lengths in amino acids of all in-frame sequences.
By design the library encodes for 101 amino-acids (dotted line). This bar graph shows the distribution of lengths for the 203 in-frame sequences.

Only for the 165 in-frame sequences of designed length the codon composition was compared to the design. The codon composition was not quite as equally distributed as designed: Guanine was significantly over-represented, whereas Adenine was under-represented for both NNK and similarly for VTH of the phi patch (see **Figure 2.29** & **Figure 2.30**).

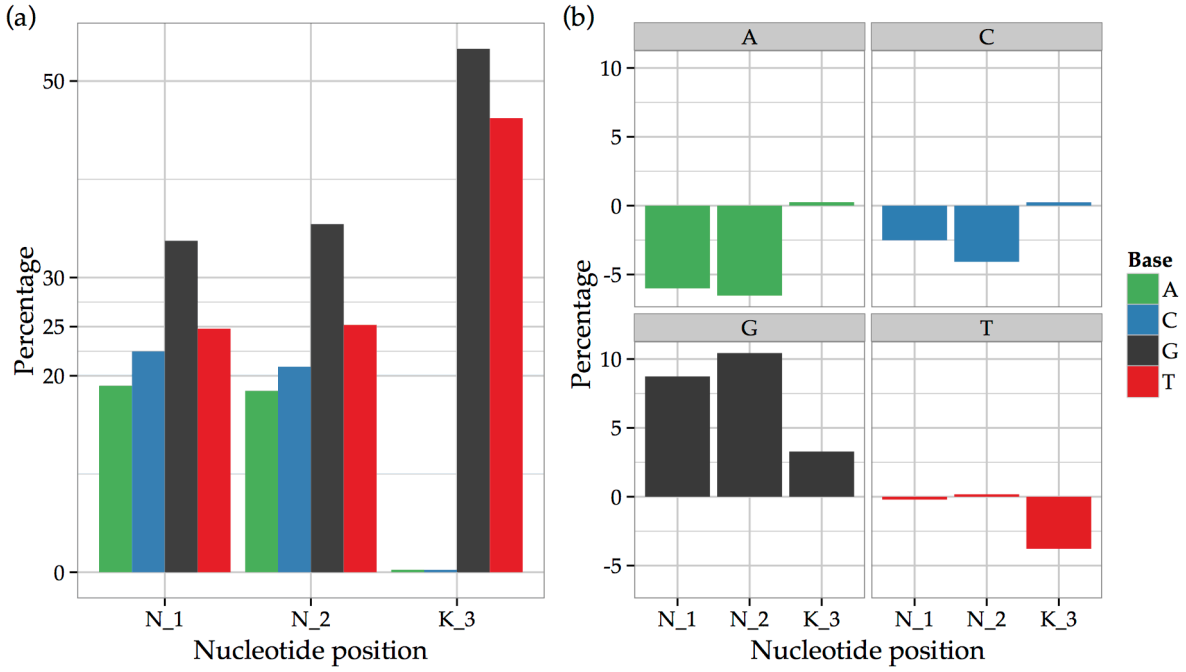


Figure 2.29: Nucleotide composition of all NNK for the in-frame sequences of designed length.
Position specific nucleotide composition for 15'840 NNK codons (the 96 NNK codons of the 165 in-frame sequences). (a) Total percentage of each nucleotide for each position. (b) Deviation from the designed, optimal composition: N=A, C, G, T each with 25% and K=G, T each with 50%.

This bias in nucleotide composition was most likely already present in the initial NNK oligonucleotides and occurred due to non-optimal mixture of bases during synthesis.

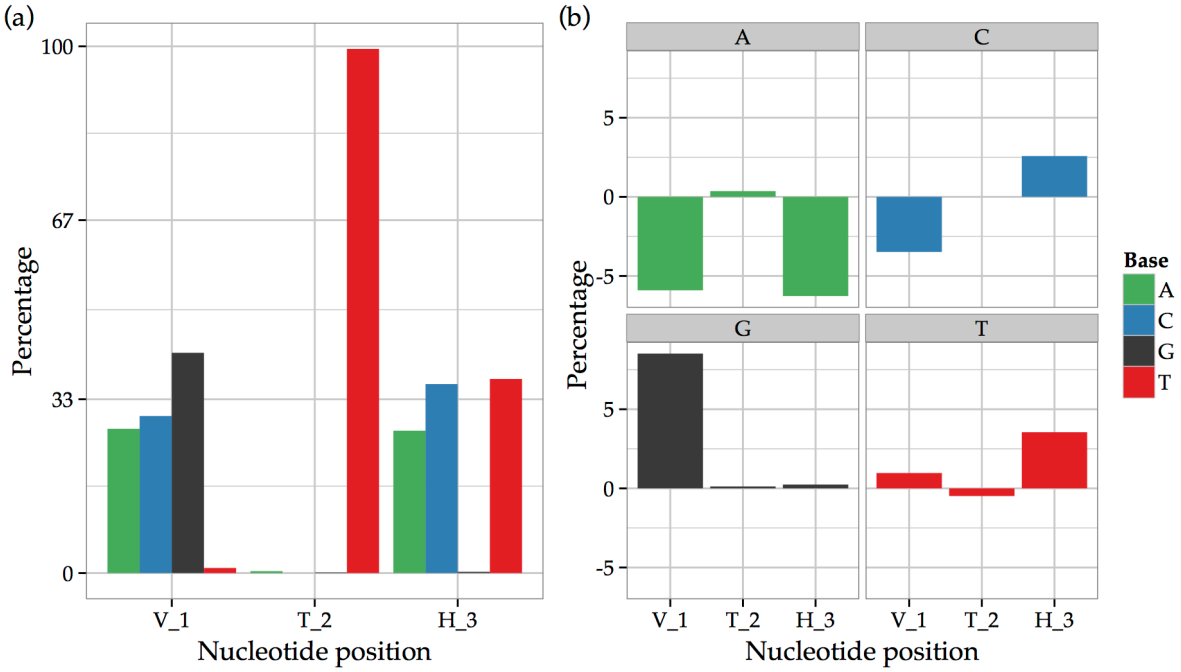


Figure 2.30: Nucleotide composition of all VTH for the in-frame sequences of designed length. Nucleotide composition for each position of 825 VTH codons (the 5 central "phi" codons for the hydrophobic patch of all 165 in-frame sequences, including the 4 constructs with point-mutations in "phi"). (a) Total percentage observed for each position and nucleotide. (b) Deviation from an optimal nucleotide composition: V=A,C,G 1/3 each and H=A,C,T 1/3 each.

The ligation overhangs showed a similar distribution, with ever higher representation of Guanine - but also keeping Adenine and Thymine in the "N" positions almost at the same percentages as observed for the total codons; inferring that the additional hybridization energy of GC pairs does *not* significantly bias the composition of ligation seams (products) (see **Figure 2.31**).

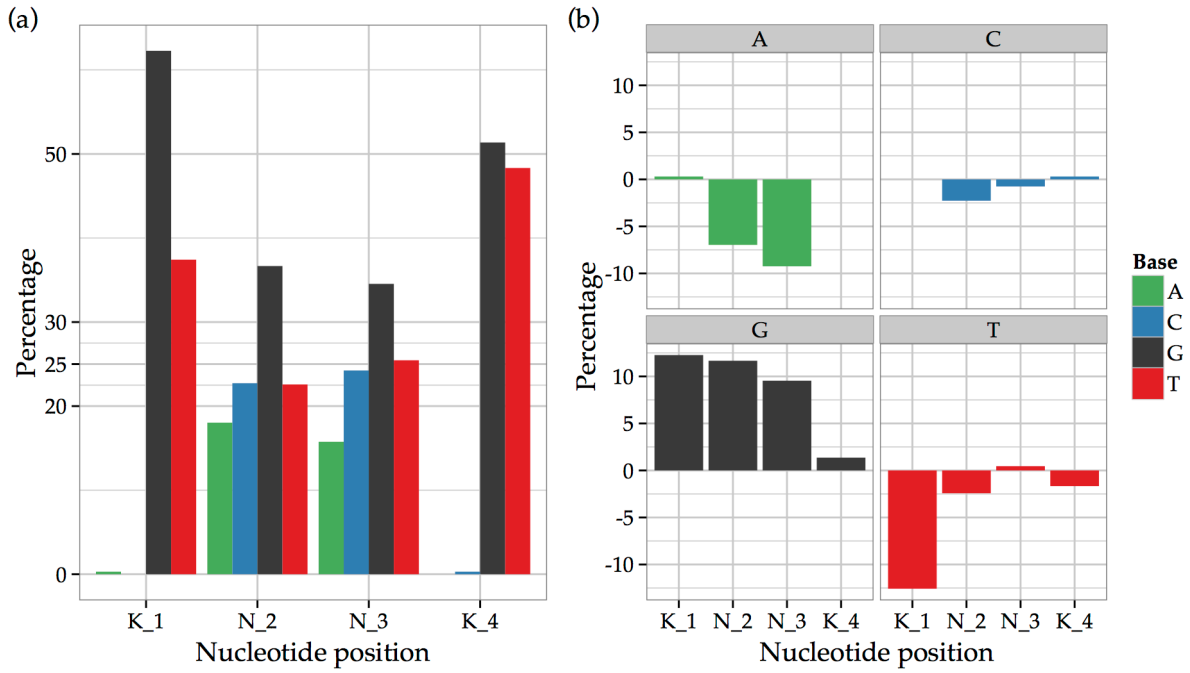


Figure 2.31: Nucleotide composition of the ligation seams for the in-frame sequences of designed length. Nucleotide composition for the 660 ligation seams (5'-KNNK overhangs) of all 165 in-frame sequences. (a) Total percentages observed for each position of the ligation seam, (b) Deviation from the optimal mixture - analogous to **Figure 2.29**.

The codon table in **Figure 2.32** shows the total number of each observed codon for in-frame sequences of the designed length of 101 encoded amino acids and color-coded the deviation to the expected number of 495 occurrences. The over-representation of Guanine and under-representation of Adenine, especially for the fist two positions, is immediately visible. The resulting effect on the levels of encoded amino acids can be seen in the dot plot of **Figure 2.33**.

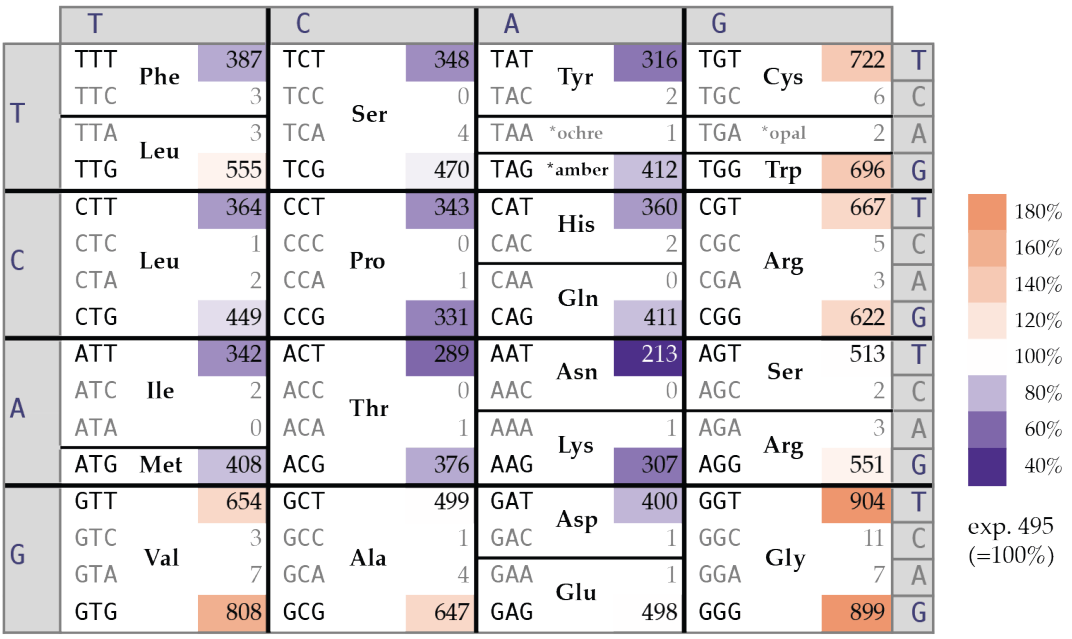


Figure 2.32: Codon-table for NNK codons of the in-frame sequences of designed length.
Codon composition for 15'840 NNK codons (the 96 NNK codons of the 165 in-frame sequences). NNK encoding 32 possible codons (printed in black). For an optimal nucleotide mixture each codon would be found 15'840/32 = 495 times. The color code shows the relative deviation to the expected occurrence.

The dot plot of encoded amino acids in **Figure 2.33** gives the ratio of encoded to the expected amino acids and thus shows the effect of the biased nucleotide mix on the level of amino acids and protein composition. The smallest amino acid Glycine (Gly/G) is strongly over-represented, whereas Asparagine (Asn/N) is extremely under-represented.

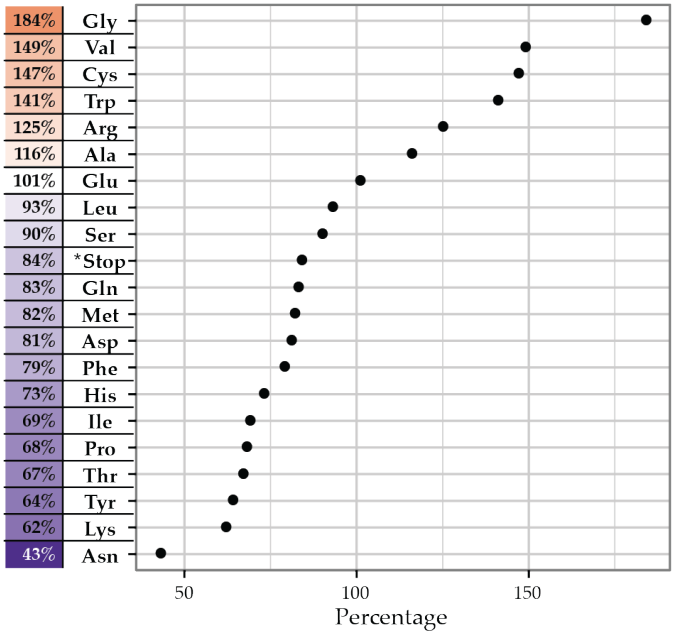


Figure 2.33: Dot plot of encoded amino-acids.
The ratio of encoded to expected amino-acid levels is plotted, showing the effect of the biased nucleotide mix on the encoded amino-acids. See also **Figure 6.3**.

2.3.5 Discussion (MOAL)

By choosing NNK codons, the probabilities for the encoded amino acids of a random protein are approximately retained, compared to all possible codons (NNN), while eliminating two of three stop-codons, making NNK a very good choice for a fully random composition of the library. The central phi module, encoding for a patch of 5 hydrophobic amino acids, was also realized as a library of five VTH codons, each position encoding Leu, Ile, and Val with the same probability.

Sanger sequencing attested a very good build quality of the final library regarding the assembly of the building blocks, the incorporation of the central patch, and the proportion of in-frame sequences.

With 62% in-frame sequences, our library has a similar or higher proportion of in-frame clones compared to previously reported random libraries [149-152].

By keeping the ligation overhangs random the library could be made seamlessly random. Importantly, only the use of signature sequences for each building block (flanking sequences like T-, G-, and B-modules and especially the flanking regions depicted in silver/gray colors in **Figure 2.27a-e**, carrying the BpiI sites) and the fusion of modules of different lengths allowed the successful completion of the library.

As ligations were not directed, having random overhangs of KNNK, it is interesting to investigate the nucleotide composition of the ligation seams for preferred base-pair configurations, e.g. if the stronger stacking interactions of G/C pairs leads to a higher representation in the observed ligation products. Comparing **Figure 2.31** and **Figure 2.29** shows that, except for the first position in KNNK, there is no significant deviation from the generally observed NNK codon composition of the library, so that the ligation seams can be deemed truly random and no bias was introduced by different hybridization energies.

The robustness of methods such as PCR, T4 DNA ligation and the high fidelity of the Type IIs restriction enzymes BsaI and BpiI allowed a construction of the library at a scale and diversity that is useful and more than sufficient for most in-vitro or in-vivo applications, i.e. mainly selections.

The experimentally obtained diversity of $\sim 10^{13}$ is of course far from the number of theoretical possible elements ($\sim 2 \times 10^{127}$), and could possibly even be fully sampled using some in-vitro technologies. Yet, each new assembly of the library would generate a completely different subset of the vast number of theoretical elements. Compared to library construction schemes involving the cloning into a vector and its transformation into bacteria [149,151,155] our library has an experimental diversity that is several orders of magnitude greater. The now worked out strategy has lead to a robust toolbox that can applied in various assembly schemes.

The quality of the oligonucleotides and accuracy of restriction digests and especially ligation made it possible to circumvent any translational (in-between selection) steps, thereby allowing the incorporation of a (central) module coding for five consecutive hydrophobic amino acids of Leu, Ile or Val, which would not have been possible with previously published methods. This module

may evidently serve as nucleation center for folded proteins by promoting the formation of a solvent-shielded hydrophobic core.

Quality control of the library by Sanger sequencing verified a high concordance to design, regarding the high proportion of 62% in-frame sequences, all containing the central phi module, and a suitable codon composition, which originates in the nucleotide mix for the wobble codons in the initial oligonucleotides. The sequencing results emphasize that an accurate nucleotide composition of the mixed positions in the oligonucleotides is pivotal, as any bias in composition not only gets carried over to the library, but also can be amplified when blocks are reused and combined. Such a bias can lead to a skewed distribution of encoded amino acids as shown in **Figure 2.33** and thereby also alter the covered sequence space.

Overall, the entirely DNA-based construction scheme (using NNK codons) and the inclusion of a central patch coding for a hydrophobic cluster (using VTH codons) facilitate an unprecedented theoretical diversity.

The circumvention of any translational steps during construction, which would have greatly reduced diversity, renders this library highly probable to cover yet unexplored areas in sequence space and makes it a powerful tool in the discovery of truly novel proteins.

2.4 Selections using the secondary structure library (SSL)

Different versions of the secondary structure library were used for selections in both the Bla and GFP setup, targeting the Tat pathway.

The SSL2.1 with its large heterogeneous size distribution was used for selections in the Bla setup, highlighting the need to eliminate short sequences present in the library and uncovering unexpected escape mechanisms when using β-lactamase as reporter.

SSL-Φ-SSL, with a central Φ (phi) module encoding a hydrophobic patch and a resolved size distribution, was used in the SF·GFP-ssrA setup for sorting fluorescent cells in FACS. One construct (in a library of 10^7 cfu) carried a point mutations leading to a stop codon before the degradation tag and dominated the selection due to its strong cytoplasmic fluorescence.

Assembly PCR of the library SSL-Φ-SSL with an fully intact SF·GFP-ssrA part created again a larger size distribution and lead to a significant shortening of the library. Although this library was not suitable for selections, it lead to the identification of a full length construct containing the hydrophobic patch, which performed well in the Bla setup but showed no signs of being well folded.

The constant shortening of the SSL with every PCR reaction, as observed again in the cloning of SSL3 without methionine after the signal sequence, lead to the omission of this library in further selections.

2.4.1 SSL2.1 in the Bla setup, selections on solid media plates

The first library selections using the Tat pathway were performed with SSL2.1 in the β-lactamase setup on selective LB-agar plates. The SSL2.1 was used in its initial state showing a large size

distribution on agarose gels. Selections were performed using both Tat signal sequences, TorA_ss or SufI_ss. With 100 µg/ml ampicillin (1 mM IPTG, 25 µg/ml Cm) rather stringent selection conditions were chosen for a first selection round. Only a small number of colonies was present on the selective LB-agar plates. These clones carried Bla-constructs that consisted mainly of short peptides. These short sequence fragments were present in the SSL2.1 in substantial amounts and with variable lengths. Additionally to the short peptides, two interesting longer constructs were observed, which employed unexpected escape mechanisms for the applied selection pressure.

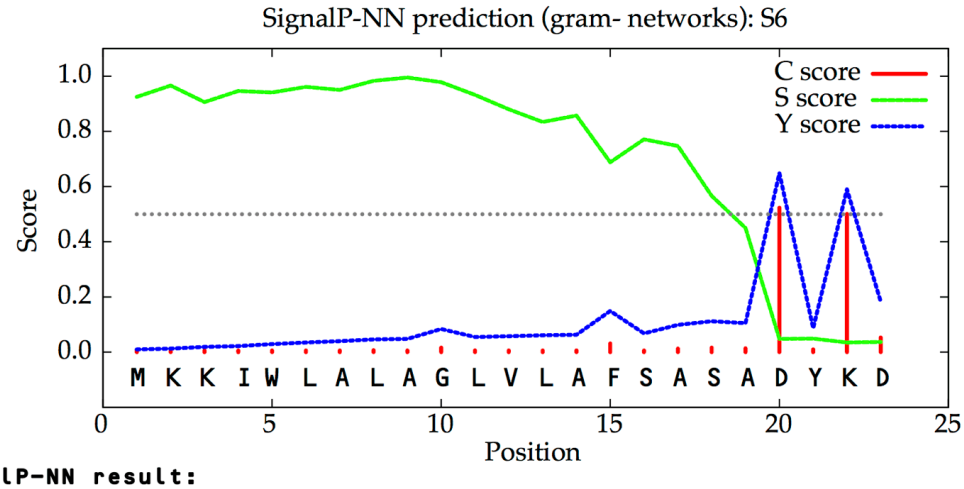
2.4.2 Novel export sequences and a constitutively active promotor

Sequencing of the longer inserts found in SSL selections with beta-lactamase revealed one escape mechanism where the DNA of the vector sequence encoding the signal peptide was rearranged. This mechanism may be similar to modifications of promotor regions observed in bacteria expressing toxic proteins [169].

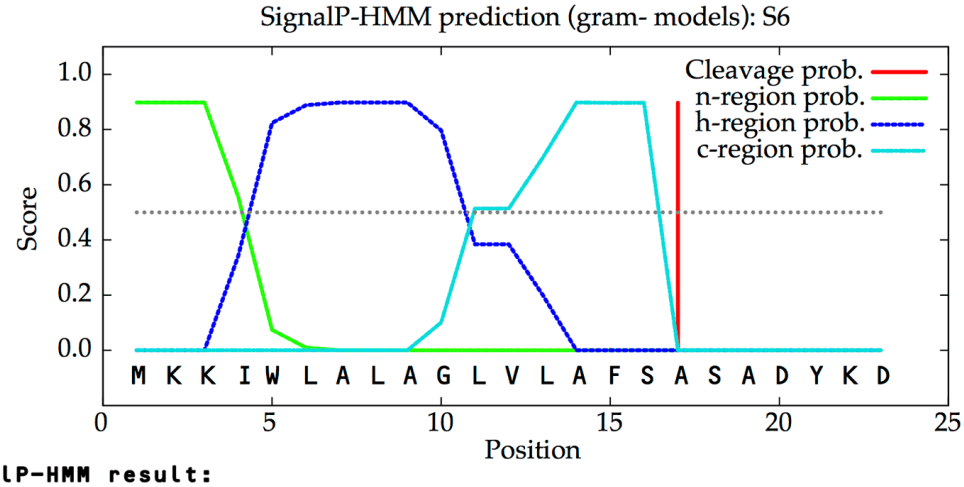
The modified sequence of this SSL clone was located before the actual cloning site used for the library and encoded a (signal-)peptide of 16 amino acids, which we termed S6. The S6 signal peptide has the amino acid sequence MKKIWLALAGLVLAFSasadykd.

Sequence alignments showed that the flanking sequence comprises the designated Lac promotor and the last 3 amino acids of the SufI signal sequence and originated from the 4L_SufI_Bla vector used for cloning.

E. coli cells deficient in *tatC* were resistant to ampicillin when expression of β-lactamase carrying this N-terminal S6 peptide was induced. The S6 peptide thus probably functions as signal sequence directing the translocation of the β-lactamase to the periplasm using a Tat-independent pathway. The bioinformatics SignalP 3.0 Server [170] identified the peptide as potential signal sequence, both using the neural network and the hidden Markov model. The more recently implemented SignalP 4.1 Server [171], returned a lower D-value (0.438) and recognized a signal sequence only when using the SignalP-noTM network and a cutoff for D at 0.34 (**Figure 2.34**).




```
SignalP-NN result:
>S6                        length = 23
# Measure  Position  Value  Cutoff  signal peptide?
  max. C     20       0.523   0.52   YES
  max. Y     20       0.649   0.33   YES
  max. S      9       0.995   0.92   YES
  mean S     1-19     0.857   0.49   YES
       D     1-19     0.753   0.44   YES
# Most likely cleavage site between pos. 19 and 20: ASA-DY
```




```
SignalP-HMM result:
>S6
Prediction: Signal peptide
Signal peptide probability: 0.898
Max cleavage site probability: 0.897 between pos. 16 and 17

# SignalP-4.1 gram- predictions (SignalP-noTM)
Cmax   pos  Ymax   pos  Smax   pos  Smean   D      ?  Dmaxcut
0.358  20   0.370  20   0.890  1    0.514   0.438  Y  0.340
```

Figure 2.34: SignalP predictions for S6 peptide.
SignalP 3.0 identifies the S6 peptide as putative signal sequence in both prediction models, short output of SignalP 4.1 for S6 with low D cutoff.

Another escape mechanism observed in the initial SSL selection consisted of multiple elements. The insert S7 contained more than 300 nucleotides and was found between the devised restriction sites (NcoI & BamHI), flanked by correct vector sequences. Interestingly, the coding sequence of the SSL2.1 insert was not in the designated reading frame of the beta-lactamase but encoded a separate, constitutively active, promotor, followed by a signal peptide, which lead to the export of the beta-lactamase to the periplasm and thus to the survival of this clone under selective

conditions. For this clone, carrying the S7 insert, resistance to ampicillin was established even without induction of the lac promotor and again independent of the Tat pathway as observed in an *E. coli* strain deficient in *tatC*.

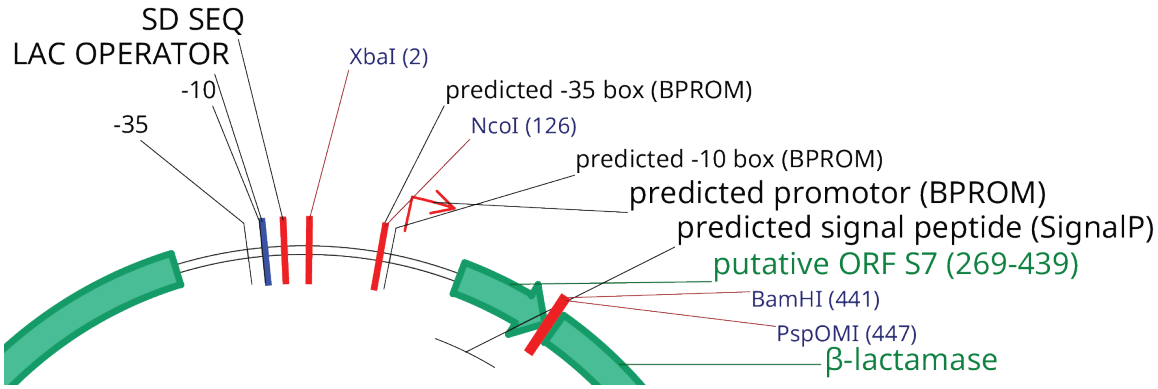


Figure 2.35: Features of the S7 insert.
Cloned via NcoI and BamHI, S7 constituted a novel promoter and Tat-independent signal sequence due to frame-shifts and/or mutations. Promotor, -35 and -10 boxes were predicted by BPROM (Prediction of bacterial promoters) [172], signal peptide was predicted by SignalP [171]. The S7 insert originated in full length from SSL2.1 and contained mainly the designed secondary structure DNA modules, albeit with a frame-shift.

2.4.3 Selections on SSL-Φ-SSL in the SF·GFP-ssrA setup using FACS

For the next selections the large size distribution of SSL2.1 was resolved and the SSL-Φ-SSL library was constructed, consisting of two SSL2.1 entities around a central Φ (phi) module that encodes a hydrophobic patch of 5 consecutive amino acids (2.2.3). To avoid survival pressure as in the case of Bla selections, the selection system was changed to TorA_ss–POI–SF·GFP-ssrA and cells were sorted in flow cytometry. To increase the dynamic range, expression was performed in the Δ*sspB* strain.

The library TorA_ss–SSL-Φ-SSL–SF·GFP-ssrA was transformed into electro-competent Δ*sspB* cells and 10^7 colony forming units (cfu) were obtained. Expression was induced with 80 µM IPTG and carried out for 12 h at 25°C.

Due to the construction scheme of SSL-Φ-SSL, the insert used for cloning the selection vector contained the library already fused to SF·GFP-ssrA. Two initial trials were performed with this library, the first trial with two sorting rounds, and the second trial with four sorting rounds in flow cytometry.

The sorted cells were recovered in medium without IPTG and then re-induced for expression and employed for further sorting. Here, the library was not re-cloned in between the sorting rounds.

In trial 1, two sorting gates were used: gate P3 contained the top 0.1% fluorescent events and the adjoining gate P2 the events comprising the top 0.7% fluorescence. In sort 1, about 95'000 events were sorted for gate P2 and ~500 events for gate P3, using the “Single Cell” precision setting. Sorted cells were recovered, followed by a second round of expression and flow cytometry sorting. In sort 2 only cells from gate P3 of sort 1 were processed due to technical difficulties with the FACS. The whole cell population showed a significantly increased fluorescence. The sorting

gates for sort 2 were adjusted to comprise events of 0.6% top fluorescence for gate P3 and 17.5% for the adjoining gate P2. Only the ~2'500 events from gate P3, originating from the ~500 events in gate P3 of sort 1, were used for further analysis.

In trial 2, four rounds of sorting were performed. The first sort was executed in high speed mode, processing 3×10^9 events, thereby oversampling the library of 10^7 cfu by 300 times. In high speed mode a threshold for the fluorescence detection is set, thereby the flow cytometer registered an average flow rate of 20/s, instead of 100'000/s to 500'000/s without this threshold. Using "Purity" precision settings, 14'000 events were sorted. As a droplet/event in this high speed mode contained a large number of cells, an estimated number of 10^5 to 10^6 cells were obtained in this sorting. Round 2 to 4 were sorted in the standard mode for 10'000 events in the sorting gate of 0.1% top fluorescence using "Single Cell" precision settings.

GFP fluorescence of single clones from the output of the final sorting round of both trials was analyzed for whole cells in flow cytometry and on solid media plates. The most fluorescent clones were screened for periplasmic localization of GFP fluorescence.

A few clones showed a high fluorescence signal for whole cells, some even higher than the positive control TrxA. Yet, none showed a significant GFP fluorescence in the periplasmic fraction, indicating that no measurable Tat-dependent translocation had occurred.

Sequencing of some of these highly fluorescent clones revealed that all carried the identical insert of 124 amino acids originating from the SSL-Φ-SSL library. The prototype of this insert was termed clone 4.24 (**Figure 2.36**).

Sequencing of the fused SF·GFP-ssrA revealed the cause of the high GFP fluorescence: all constructs carried a point mutation in the C-terminal tail of SF·GFP before the ssrA degradation tag. This point mutation created a stop codon (GAA → TAA) and resulted in a construct with high cytoplasmic GFP fluorescence due to the absence of a degradation tag (**Figure 2.36**).

Although the different clones were all identical in the sequenced region including the Shine-Dalgarno sequence, they showed different whole-cell fluorescence levels. Two variants were re-transformed and maintained their fluorescence profiles, suggesting that further mutations in their plasmids may be responsible for the differences in expression.

The SSL-Φ-SSL insert of clone 4.24 was also cloned into the Bla setup with both Tat signal sequences, TorA_ss and SufI_ss. Assayed in a droplet dilution series on solid media plates, the construct performed comparable to or worse than the negative controls.

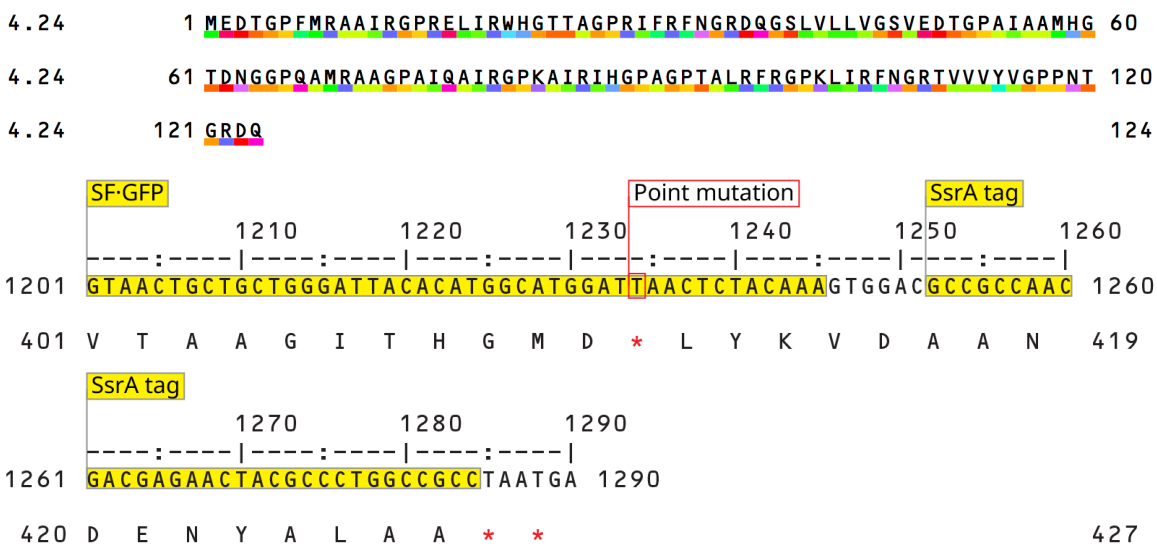


Figure 2.36: Sequence of the 124 amino acid insert (clone 4.24) and site of the mutation creating a stop-codon before the degradation tag. Underline coloring of the amino acids analogous to Taylor's aminochromography [173].

2.4.4 Screening for fluorescent phenotypes on solid media plates

Screening of bacterial colonies for whole cell fluorescence can also be performed on LB-agar plates using a light source with appropriate spectral properties and a longpass or bandpass filter for the respective fluorophore.

Single clones from the GFP-selections using flow cytometry sorting on SSL-Φ-SSL were analyzed on solid media plates with IPTG. Plates containing dispersed colonies were imaged in a Fujifilm Image Reader LAS-3000 using blue light (460 nm EPI) for excitation and a Y515AttoPhos filter to detect GFP fluorescence. Colonies showing fluorescence above a threshold were transferred to an extra solid media plate to compare their fluorescence to cells expressing control constructs, spotted on the same plate (**Figure 2.37**).

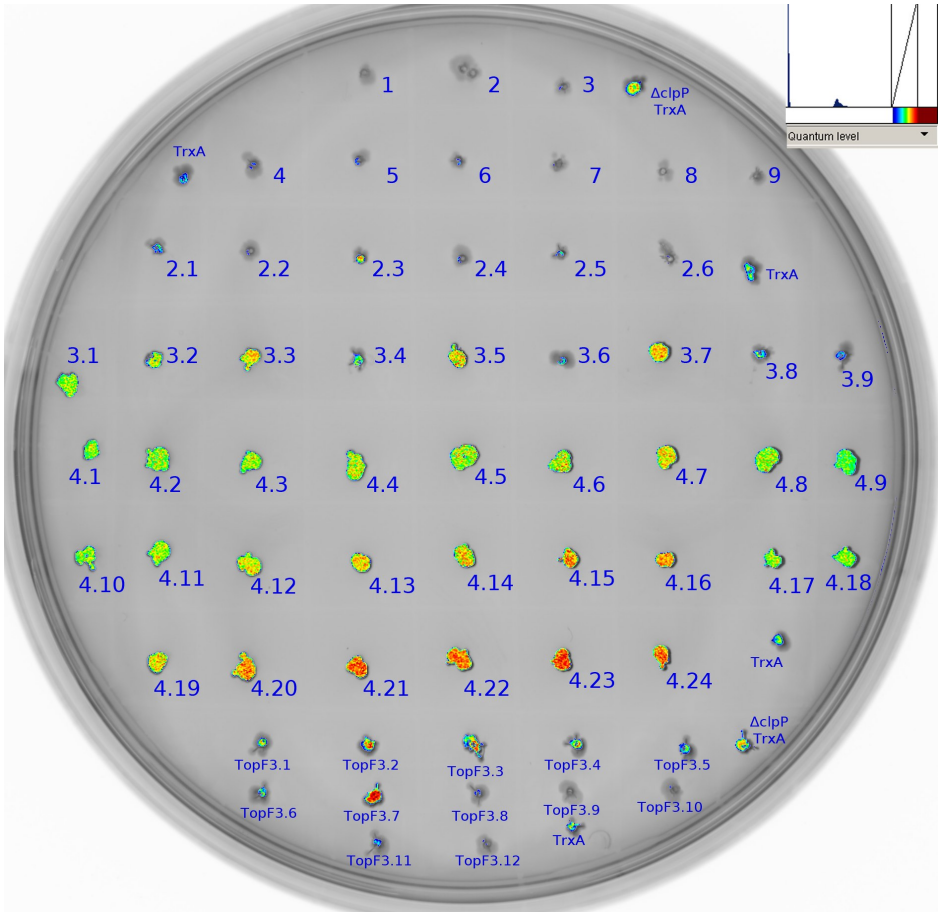


Figure 2.37: Fujifilm Image Reader LAS-3000 photography of a solid media plate with single clones spotted.
Image was overlaid with GFP fluorescence signals represented in a rainbow color gradient (see inset in the upper right corner).

Most of the highly fluorescent clones expressing the winner construct of the selections on SSL-Φ-SSL, carrying a stop codon in the C-terminal tail of SF·GFP before the degradation tag, were identified on solid media plates. Fluorescence levels from flow cytometry and from colonies analyzed in the Image Reader LAS-3000 correlated generally. Flow cytometry allowed a more precise ranking, whereas fluorescence scans of whole plates could be performed with higher throughput. Yet, both methods record the GFP fluorescence of whole cells and give no indication if the GFP fusion constructs are localized in the periplasm.

To acquire a readout that correlates with Tat-dependent translocation and thereby is coupled to the folding properties of a protein, an effective screening method must verify the localization of the GFP fusion proteins in the periplasm. Fractionation by cold osmotic shock and the measurement of the periplasmic fraction was found to be the most conclusive screening method and was therefore used preferentially in further experiments.

2.4.5 Selections on re-cloned SSL-Φ-SSL, analysis of A11

Selections on SSL-Φ-SSL were dominated by one construct with inactive degradation tag. The respective stop-codon mutation was found to be present in the SSL-Φ-SSL–SF·GFP-ssrA library insert used for cloning. Yet, we did not want to include the GFP sequence in the cloning vector to

avoid false positives due to re-ligated vector, possibly leading to a TorA_ss–SF·GFP-ssrA construct with high periplasmic fluorescence. The part of SF·GFP-ssrA was re-amplified by PCR from an intact sequence without stop-codon and used for re-cloning the library.

The library SSL-Φ-SSL was cut at the BamHI site from the full insert SSL-Φ-SSL–SF·GFP-ssrA. On an non-denaturing agarose gel the cut DNA of the library ran as a sharp band at ~400 bp, indicating that SSL-Φ-SSL in this state still encoded for more than 100 amino acids. To fuse the library SSL-Φ-SSL to the freshly prepared SF·GFP-ssrA from a template without stop-codon, different strategies were tested. Ligation, ligation-independent cloning (LIC), and assembly PCR of the two fragments were compared regarding the yield of fused product SSL-Φ-SSL–SF·GFP-ssrA. The assembly PCR resulted in the highest yield of fused product, followed by LIC. Ligation of the cut fragments showed no visible band on agarose gel at the expected size of a fused product.

However, any amplification of SSL-Φ-SSL with flanking primers, both for LIC and the assembly PCR with SF·GFP-ssrA, resulted in a significant shortening of the library as well as a larger size distribution, observed as broadening and blurring of the respective band on agarose gels.

The product of the assembly PCR was finally used for re-cloning, followed by transformation and expression in Δ*sspB* cells. In a single sort using flow cytometry 3×10^5 event were collected for the top 1% in GFP-fluorescence, using "Yield" precision settings.

Analysis of a dozen single clones from the sorting showed mainly short peptide inserts, leading to a good Tat-dependent export rate and high periplasmic signal. One construct, A11, carried a full length insert, including the hydrophobic patch, and had a significant periplasmic fluorescence signal at ~60% of the positive control 4L_TrxA.

The SSL-Φ-SSL insert of A11 was cloned into a freshly prepared TorA_ss–SF·GFP-ssrA vector and re-assayed for periplasmic fluorescence. It was further analyzed in the Bla setup, using TorA_ss, and cloned into an expression vector to assess the capability of expressing and purifying the A11 protein without fusion partner.

The re-cloned TorA_ss–A11–SF·GFP-ssrA showed no significant periplasmic fluorescence, indicating that the initial clone had probably been a double transformant of the A11 insert and a short peptide insert.

Yet, A11 performed rather well in the Bla setup, where it showed better export than all the negative controls (**Figure 2.38**). This was the first example of a polypeptide with an extended hydrophobic patch that performed reasonably well in the Bla assay on solid media plates.

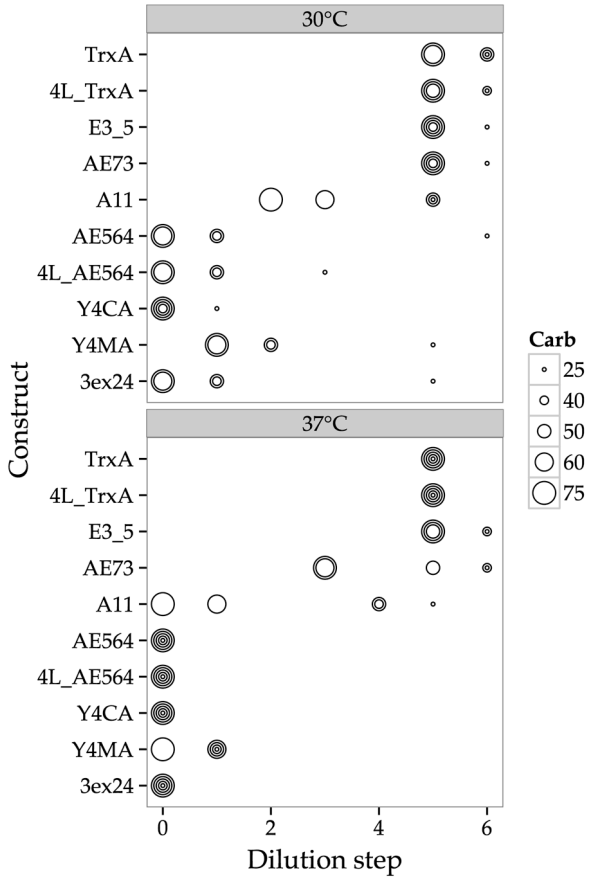


Figure 2.38: Bla dilution assay (1:10) of construct A11 in comparison to other test proteins. All constructs with TorA_ss, assays were carried out on solid media plates with 150 µM IPTG and 5 different carbenicillin concentrations at 30°C and 37°C.

To test cytoplasmic expression without fusion proteins, A11 was cloned into the expression vector pEc_HP-lacIq-T5-rgs_his6-SacB (eLIC_051). The first expression test was performed at 30°C for 16 h in *E. coli* XL-1 blue cells using 2YT medium and 1 mM IPTG induction at OD_{600} = 0.5. No band corresponding to MRGSHHHHHH-A11 could be detected in Coomassie Brilliant Blue stained SDS-PAGE of whole cell lysates, nor in the elutions from Ni-NTA IMAC of the soluble protein fraction. Additionally, DH5α, JM83, and TOP10F' were transformed with the A11 expression vector. An anti-RGSHHHH western-blot of samples after 3.5 h expression with 0.5 mM IPTG at 30°C or 37°C did not show any detectable bands for the expression of A11. These results suggested that A11 was most probably not a well-folded protein, difficult to express in *E. coli* or very prone to degradation, and thus of little use in the search for truly novel proteins.

2.4.6 Selections using SSL3

During the construction of the library SSL3 a shortening of the library was once more observed when using flanking primers for amplification (see 2.2.4). In the assembly PCR with SF·GFP a further reduction of the library length occurred. Analysis of single clones after one round of sorting in the flow cytometer revealed that most constructs had lost the hydrophobic patch, consisted of only few modules and encoded for short peptides.

Due to the persistent shortening problem in PCRs, the secondary structure library was not used for further experiments. At this stage it became clear that a fresh assembly of an SSL3 type library would be needed. This would include the incorporation of the hydrophobic patch module in-between two SSL entities and more importantly the challenge to find and establish a revised set of flanking primer sequences, which could perhaps prevent the continuous shortening of the library in PCR reactions.

2.5 Selections on the MOAL

As the SSL, with its high intra-module sequence similarity, has been prone to produce significant unfavorable side-products with each round of PCR amplification, and more importantly covers only restricted spots in sequence space, a new library was devised for selections towards truly novel folded proteins.

The MOAL, a fully random NNK library with a central Φ (phi) module of five VTH codons encoding 5 consecutive hydrophobic amino acids (Leu, Ile, Val), was therefore assembled (2.3) and used in both Tat setups for selections.

Selections were performed in *E. coli* DH5α by alternating between Bla and GFP setup and in each round the full-length library was purified by gel extraction and ligated into freshly prepared vector.

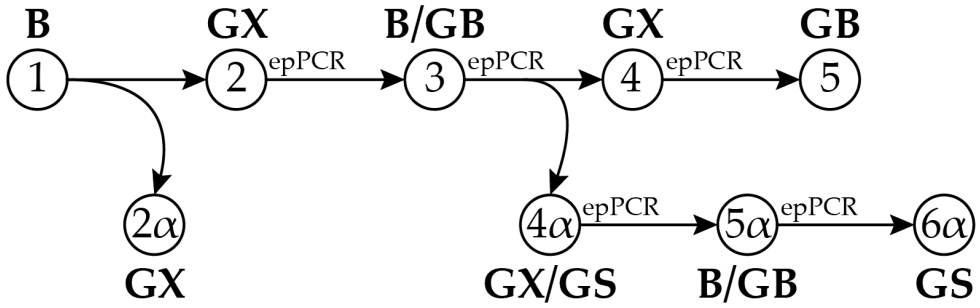


Figure 2.39: Overview of selection rounds performed with the random library.

Round	Reporter	Sets	Temp./°C	Time/h	IPTG/µM	carbenicillin /µg/ml
1	B	2	30, 30	3.5, 4.5	100	20
2α	GX	1	30	4	150	
2	GX	1	23	11	90	
3	B/GB	2	30, 33	3.5, 2.5	100	40
4α	GX/GS	1	30	3.5	85	
4	GX	1	28	12	60	
5α	B/GB	2	30	3.5	90	45
5	GB	2	29	4.5	75	50, 70
6α	GS	2	23, 28	12, 11	110	

epPCR: error-prone PCR on output of selection round; B: Bla setup; GB: SF·GFP-Bla setup used in antibiotic resistance selections; GX: SF·GFP-SclpX setup; GS: SF·GFP-ssrA setup.

2.5.1 Round 1: Bla selections by filtration

The completely in vitro assembled MOAL was cloned via the flanking restriction sites BamHI and PspOMI into the ssTorA_Bla vector and transformed by electroporation to obtain transformation numbers of 10^9 colony forming units (cfu).

In the first round, Bla selections for β-lactam antibiotic resistance were performed at 30°C in 20 ml liquid medium (20 µg/ml carbenicillin, 100 µM IPTG) by 5 µm filtration. As the experimental diversity for the first selection was 10^9 and 1 ml of $OD_{600} = 1$ corresponds to ~0.5×10^9 cells, 20 ml medium were inoculated with *E. coli* DH5α cells to a starting OD_{600} of 0.15, which corresponds to 1.5×10^9 cells.

Expression was induced with 100 µM IPTG at $OD_{600} = 0.6$ and after 3.5 h shaking at 30°C the cell suspension of OD_{600} ~4.2 was passed through a 5 µm filter; after filtration the OD_{600} was ~0.03.

Presumably, most cells contained constructs that are not compatible with Tat-dependent translocation. The fused β-lactamase of these constructs was not exported to the periplasm and the cells expressing these constructs were not resistant to β-lactam antibiotics. Their cell division was blocked by the antibiotic and these bacteria grew too large to pass the 5 µm filter.

Only cells that express constructs where the protein encoded by the library can be exported by the Tat pathway, as TorA_ss–POI–Bla fusion, would contain periplasmic beta-lactamase and will thereby be able to undergo cell division and maintain a size that can pass the 5 µm filter. If we assume a doubling time of 1 h for dividing cells under these selective conditions, the cells that passed the filter after 3.5 h expression had undergone on average $2^{3.5} \approx 11.3$ divisions. Thus, the OD_{600} of the filtered cells of 0.03 corresponds to an OD_{600} of $0.03/11.3 \approx 0.0027$ during induction, which is 0.45% of cells present at the time point of induction. In 20 ml of $OD_{600} = 0.6$ there were 6×10^9 cells upon induction and by selecting 0.45% the experimental diversity is reduced to 2.7×10^7 after the filtration.

The cells were sedimented after filtration and resuspended in non-selective medium for recovery. A second analogous series of induction and filtration was performed on the recovered cells. This time, the filtration was performed 4.5 h after induction at an OD_{600} of 9.3, and the OD_{600} after filtration was 0.4.

The library was amplified from plasmid DNA obtained after selection using flanking primers (fw_NNK_t & rv_NNKb, see also 2.3.4). Gel extraction of the full-length library guaranteed an efficient elimination of short sequences, which could otherwise dominate selections.

2.5.2 Round 2α and round 2: SF·GFP-SclpX selections by FACS

The second selection round on the MOAL was performed in the GFP setup, using the TorA_ss–SF·GFP–SclpX format, carrying the slower degradation tag SclpX (2.1.9).

Round 2α lead to a new type of false positive signal, where the constructs stay fluorescent in the cytoplasm but are no longer accessible to the ClpXP degradation system. This could be attributed to three major points: the IPTG concentration, the expression temperature and the stringent setting of the FACS gate. Expression was induced with a final concentration of 150 µM IPTG and carried out for 4 h at 30°C.

With increasing expression rates due to higher IPTG concentrations or higher temperatures, the Tat pathway may no longer be able to efficiently translocate the heterologous proteins. If the Tat-

targeted proteins are not well folded, they should not be translocated and a high expression rate may overload the ClpXP-dependent degradation capacities, leading to accumulation of cytoplasmic fluorescence (compare **Figure 2.14b**).

The FACS sorting gate was further set for stringent sorting, which included only the highest 0.4% in fluorescence of the sample and was not restricted regarding the maximal fluorescence, going orders of magnitude beyond the maximal fluorescence signal of the positive control, where only constructs aggregating fluorescently in the cytoplasm or periplasmic constructs consisting of a short peptide fused to SF·GFP would locate. Some constructs from round 2α are described in the next section (2.5.3).

After identification of the critical parameters round 2 was carried out using a library with an experimental diversity of 7×10^7 cfu. Expression was driven by 90 µM IPTG at 23°C for 11 h. In order to achieve a good sampling of the experimental diversity, more 10^9 events were processed at the flow cytometer AriaIII and sorted for the top 1% in GFP-fluorescence (FITC channel) to obtain 3×10^6 sorted events (sorting mask: yield).

This less stringent sorting gate was set to include events up to a maximal fluorescence of the positive control TorA_ss–TrxA–SF·GFP-SclpX and should be better suited to obtain constructs compatible with the Tat folding check, which show certain characteristics of folded proteins and can help to further narrow down the sequence spots where truly novel, well folded proteins can be found.

Any events with higher fluorescence than (98% of) the positive control were discarded, as they most likely originate from short peptides fused to GFP or fluorescent constructs in the cytoplasm, which are aggregating or inaccessible to ClpXP degradation.

After recovery of the sorted cells the plasmid DNA was obtained and the library was amplified by error-prone PCR to increase diversity for round 3.

2.5.3 Characterization of non-translocated false positives from round 2α

Analytical flow cytometry was performed for 48 single clones from sort 2α. In colony-PCR all 48 clones showed bands corresponding to full length inserts. Expression for flow cytometry was initially performed with the similar parameters as for round 2α, 150 µM IPTG and 4 h at 30°C.

The 28 most fluorescent clones were expressed again at two different conditions, one similar to before with 150 µM IPTG and 4 h at 30°C and one at a reduced rate with 90 µM IPTG and 6 h at 26°C. Periplasmic fractions and whole cell fluorescence was measured at a plate reader. Eleven of the constructs showing the highest whole-cell fluorescence in flowcytometry and on the plate reader were sequenced, all contained the hydrophobic patch, three were without stop-codon and one introduced a frame-shift before the GFP.

None of the 28 tested construct showed significant fluorescence in the periplasm when expressed at 26°C, inferring that no (Tat-dependent) translocation occurred.

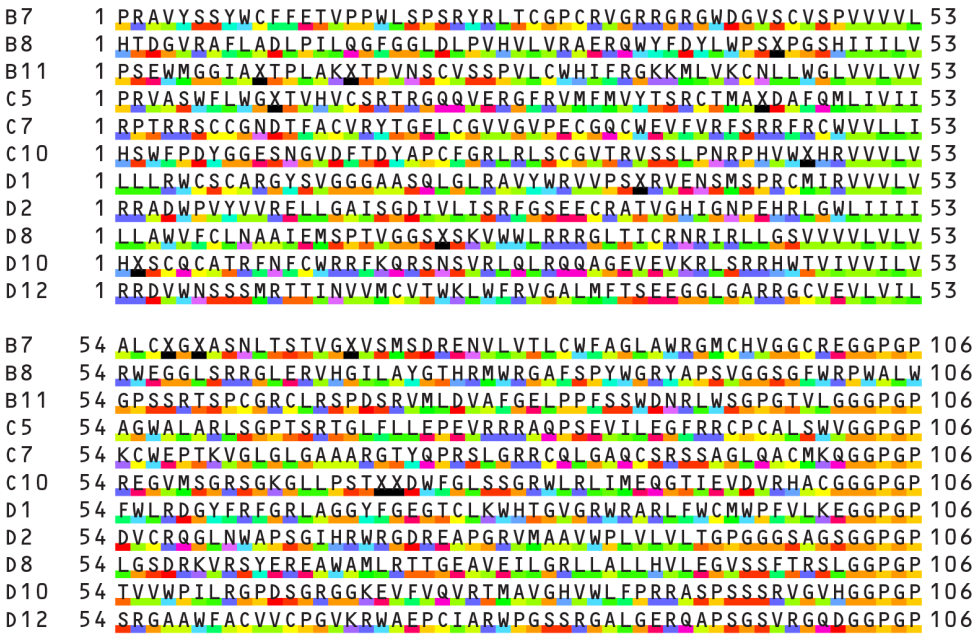


Figure 2.40: Alignment of amino acid sequences of some of the constructs from selection round 2α having higher whole-cell fluorescence levels than the positive control TorA_ss–TrxA–SF·GFP-SclpX. All constructs were of full length (101 aa) and contained the hydrophobic patch at position 49 to 53. B8 contained a frame-shift before SF·GFP, amino acids as single letter codes, X = stop codon, underline coloring of the amino acids analogous to Taylor's aminochromography [173].

To test the aptness of the Bla system for counter selection against constructs prone to oligomerization and aggregation, seven of the most fluorescent clones from this screening were re-cloned into the ssTorA_Bla vector. Additionally, a new fusion reporter SF·GFP-Bla was created, to assess the role of the adjoining protein on the aggregation and export propensity of the protein of interest. If a certain protein is prone to aggregation as SF·GFP fusion, it may behave differently when fused to the β-lactamase and thereby get carried over into the next rounds. The SF·GFP-Bla fusion was designed to keep the direct environment of the POI as constant as possible and not to change its aggregation properties by fusing it to a different protein. Thus, proteins that aggregate only when fused directly to SF·GFP could also be selected against in a β-lactam antibiotic resistance selections.

In a dilution series on selective LB-agar plates the export rates of these 7 constructs were characterized in both formats, as Bla fusion and as SF·GFP-Bla fusion. Different stringencies were tested, 30°C or 37°C with 50 µg/ml or 100 µg/ml carbenicillin on the plates. LB-agar pates with 25 Chloramphenicol and 100 µM IPTG were used, dilution steps were 1:16. As only 4 dilution steps were printed, the number of colonies or the area of the droplet covered by colonies for the highest dilution step with visible cell growth was taken into account and translated into an decimal place to allow a more detailed comparison of the used constructs (**Figure 2.41**). The construct TorA_ss–B7–Bla conferred resistance to ampicillin even when expressed in the Δ*tatC* strain.

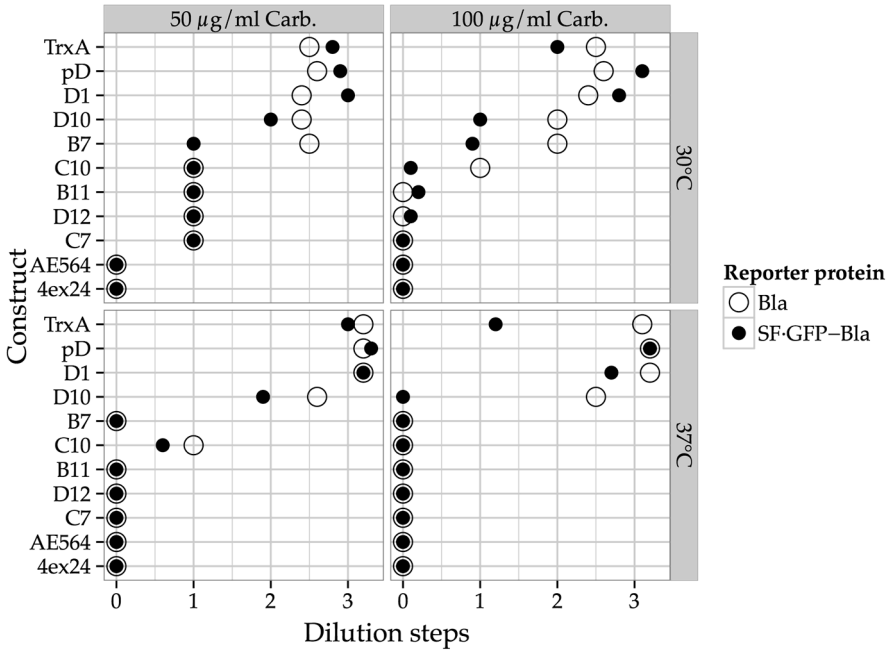


Figure 2.41: Bla and SF·GFP-Bla dilution assay (1:16) of false positive constructs from round 2α with TorA_ss on solid media plates with 100 µM IPTG and 2 different carbenicillin concentrations at 30°C and 37°C.

The two constructs without stop codon, D12 and C7, and the construct with the highest fluorescence as SF·GFP-SclpX fusion with TorA_ss, B11, were further examined as SF·GFP-ssrA and S65T·GFP-SclpX fusions. There was no measurable periplasmic fluorescence for these constructs in any format. However, very high fluorescence signals for the cytoplasmic fraction could be recorded for the SF·GFP fusions, both with the weakened SclpX degradation tag and the strong ssrA tag.

Using the wild-type like GFP with lower stability and loss of fluorescence upon aggregation, the three constructs were also assayed as S65T·GFP-SclpX fusions. The fluorescence signals of whole cells and the periplasmic fraction were low and close to background, similar as for the controls TrxA, gpD, E3_5, AE564, and AE73. However, all three constructs, B11, C7, and D12, showed significant fluorescence over background in the cytoplasmic fraction, even as S65T·GFP-SclpX fusions (**Figure 2.42**). This cytoplasmic signal was probably too low to be detectable in whole cell recordings using the wild-type like GFP.

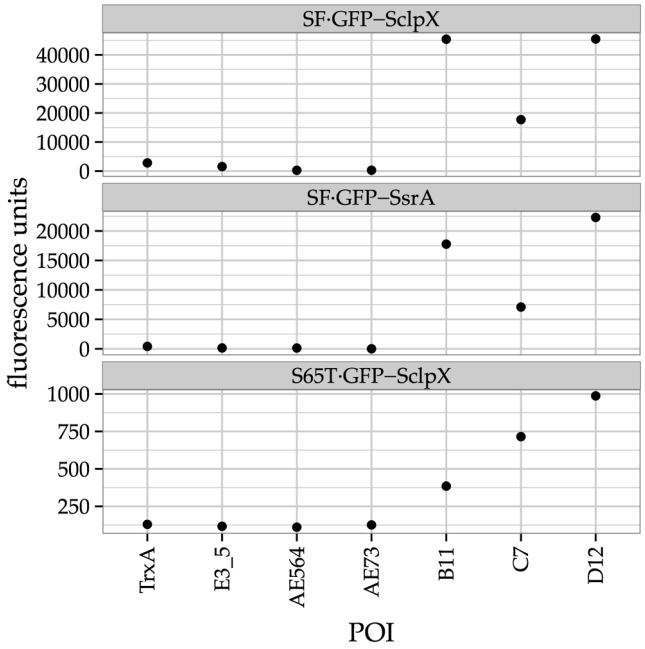


Figure 2.42: Fluorescence recordings of DH5α spheroblasts after cold osmotic shock preparation of false positive constructs from selection round 2α
B11, C7, and D12 showed measurable localization of fluorescence only in the cytoplasm, even with the strong degradation tag ssrA and as S65T·GFP-SclpX fusions; all constructs with TorA_ss.

The significant GFP fluorescence signal of these false positive constructs did not locate in the periplasm due to successful translocation via the Tat pathway, but was found only in the cytoplasmic fraction. Despite carrying a degradation tag, these constructs were able to escape degradation by ClpXP and, moreover, could not be eliminated in Bla-based setups.

Such a combination of attributes showed a major limitation of the employed setups and possibly of any selections based on Tat translocation.

2.5.4 Round 3: Bla & SF·GFP-Bla selections by filtration

In the third round, selections for β-lactam antibiotic resistance were again performed by 5 µm filtration. In addition to the ssTorA_Bla vector, the TorA_ss–SF·GFP–Bla vector was used in parallel to enhance the probability for effective counter selections against aggregating constructs.

The experimental diversity after transformations in DH5α was 10^8 cfu for the SF·GFP-Bla format and 2.4×10^8 cfu for ssTorA_Bla.

Selection conditions were similar to round 1, with the key difference of selection pressure by β-lactam antibiotic concentration, which was 40 µg/ml carbenicillin in this round.

Induction was performed by 100 µM IPTG and a first series of expression at 30°C was carried out for 3.5 h, followed by a recovery after filtration in selective medium at 22°C and another selection series at 33°C for 2.5 h.

The library in the SF·GFP-Bla format showed significantly lower OD_{600} after filtration, possibly indicating a more stringent selection. After recovery of the second filtration the plasmid DNA was

isolated for both formats individually, and for both the ssTorA_Bla and TorA_ss–SF·GFP-Bla format the library was separately amplified by error-prone PCR.

2.5.5 Round 4α, 5α, and 6α

In round 4α the error-prone PCR on the output of both formats of round 3 was used as insert. Flow cytometry sorting was performed in the GFP setup using both the TorA_ss–SF·GFP-SclpX format with the slower degradation tag SclpX, and the TorA_ss–SF·GFP-ssrA format with the fast ssrA degradation tag.

The experimental diversity after transformations was 3×10^7 cfu for the SF·GFP-SclpX setup and 10^7 cfu for the SF·GFP-ssrA setup. Induction was done with 85 µM IPTG and expressions went for 3.5 h at 30°C.

Due to a miscommunication, the FACS gate was set to a narrow band containing the top 2% in GFP fluorescence of the library in each format, with the gate cutoff being at 99% of the maximum fluorescence of the respective library, not the positive control. For the library in the SF·GFP-ssrA format 6.6×10^5 events were sorted and 2.2×10^6 events for the SF·GFP-SclpX format.

After recovery of the sorted cells the plasmid DNA was extracted and the library was amplified separately by error-prone PCR for both the SclpX and ssrA format. For the insert of round 5α the output of both formats of round 4α was pooled.

Round 5α consisted again of β-lactam antibiotic resistance selections by filtration with 5 µM filters. Both vector formats, the Bla and the SF·GFP-Bla format, were used and the diversity after transformations was 2×10^7 cfu for the Bla format and 3×10^7 cfu for SF·GFP-Bla. 90 µM IPTG was used for induction, 45 µg/ml carbenicillin and a temperature of 30°C as selective pressure. Filtrations were performed 3.5 h after induction followed by sedimentation and resuspension in selective medium for an identical second series of expression and filtration.

Again, the SF·GFP-Bla format showed significantly lower OD_{600} after filtration, compared to the Bla format. Recovery and preparation of plasmid DNA was performed individually for both formats, the Bla and the SF·GFP-Bla format.

The library was amplified by error-prone PCR on both formats separately.

The insert for round 6α was composed of the pooled output after epPCR from round 5α. Only the SF·GFP-ssrA format with the strong ssrA degradation tag was used for sorting. The diversity after transformation was 10^8 cfu. Expression was induced with 110 µM IPTG and carried out for 12 h at 23°C. The FACS gate was set with the desired parameters, beginning at the top 2% in fluorescence of the library and ending at an upper limit including 95% of the fluorescence signal of the positive control. More than 3×10^8 events were processed and 1.7×10^6 events were sorted. Cells were recovered and used for a second expression with 90 µM IPTG and at 28°C for 11 h. A final sorting was performed with similar settings as before.

2.5.6 Characterization of the dominant construct after selection round 6α

Sequencing of 19 clones after the last selection showed that the majority (68%) of constructs originated from one sequence coding for 96 amino acids, containing the central hydrophobic patch (**Figure 2.43**). These constructs, however, showed no periplasmic fluorescence. All fluorescence was located in the cytoplasm. The constructs obtained after round 6α probably had acquired similar traits as the false positives from the stringent round 2α, where Bla selections could not fully eliminate constructs that were not translocated via the Tat pathway as GFP fusion but evaded cytoplasmic degradation despite carrying a specific degradation tag.

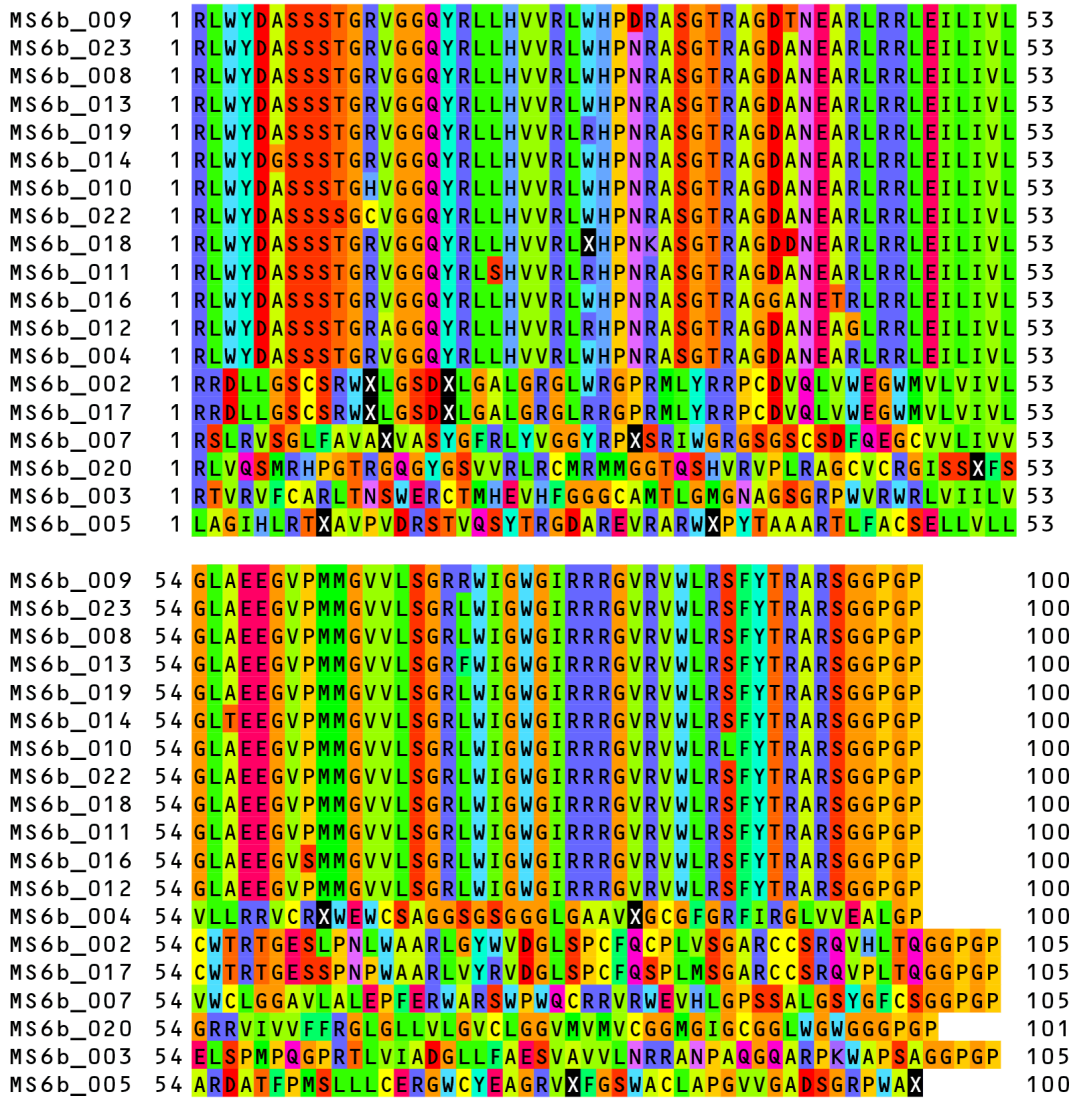


Figure 2.43: Sequencing results after round 6α
The output was dominated by variants of one construct of 96 amino acids, containing the central hydrophobic patch. Background coloring of the amino acids analogous to Taylor's aminochromography [173].

2.5.7 Single clone analysis from round 3, 4α, 5α, and 6α

The output of round 6α second sort, 6α first sort, 5α, 4α, and round 3 was cloned into the TorA_ss–SF·GFP-SclpX format. 96 single clones of each transformation were picked and expressed in small scale with 100 µM IPTG at 30°C for 4 h. Periplasmic fractions were prepared

and measured in the plate reader; no fluorescence significantly above background was observed. Sequencing of several clones from each round showed that the dominant sequence after round 6α was also the most abundant sequence in the output of round 5α. Sequences from round 4α showed a comparatively high proportion of frame shifts.

In 62 sequenced constructs after round 3 the dominant sequence of round 5α and 6α was not found. Of these 62 sequences, 45 were without stop codon (73%) and only 4 had a frame shift (6%)

2.5.8 Round 4 and 5

The sequencing of constructs after round 3 and 4α to 6α revealed that the library consisted of mainly one dominant construct from round 4α on, possibly due to the erroneous setting of the sorting gate in round 4α. However, this dominant construct showed no Tat-dependent periplasmic localization.

A new selection round 4 was performed, using only the format with the weaker degradation tag SF·GFP-SclpX. The same library as in round 4α was used, having a diversity of 3×10^7 cfu. Expression was induced with 60 µM IPTG and carried out for 12 h at 28°C.

Here, the sorting gate was set similar to round 6α, including the top 3.5% in fluorescence of the library. 2×10^8 events were processed and 6.2×10^6 events sorted. After recovery the plasmid DNA was used for error-prone amplification of the library.

Lastly, selections for β-lactam antibiotic resistance by 5 µm filtration were performed in round 5, using only the SF·GFP-Bla format. After transformations, the diversity was 10^7 cfu. Expression was induced with 75 µM IPTG and carried out at 29°C for 4.5 h. Two different carbenicillin concentrations were used, 70 µg/ml for the selection with high stringency and 50 µg/ml for medium stringency. After filtration the cells were recovered and the plasmid DNA was isolated, containing the output of the library after five alternating selection rounds.

Sequencing of 21 clones after round 5 resulted in six short sequences (encoding inserts of less than 26 amino acids), two frame-shifts, and seven clones originating from one sequence encoding an insert of 95 amino acids, which dominated the full length sequences (**Figure 2.44**). Only two sequences were without stop codons, G03_3781 and G12_3781. Most of these full-length sequences were found again in the large-scale screening of the output from round 5, including the dominant construct without stop codon. None of the longer constructs identified here showed significant periplasmic fluorescence. They probably were selected due to similar characteristics as the false positives from round 2α and the winner of round 6, having significant cytoplasmic fluorescence and getting exported as Bla fusions.

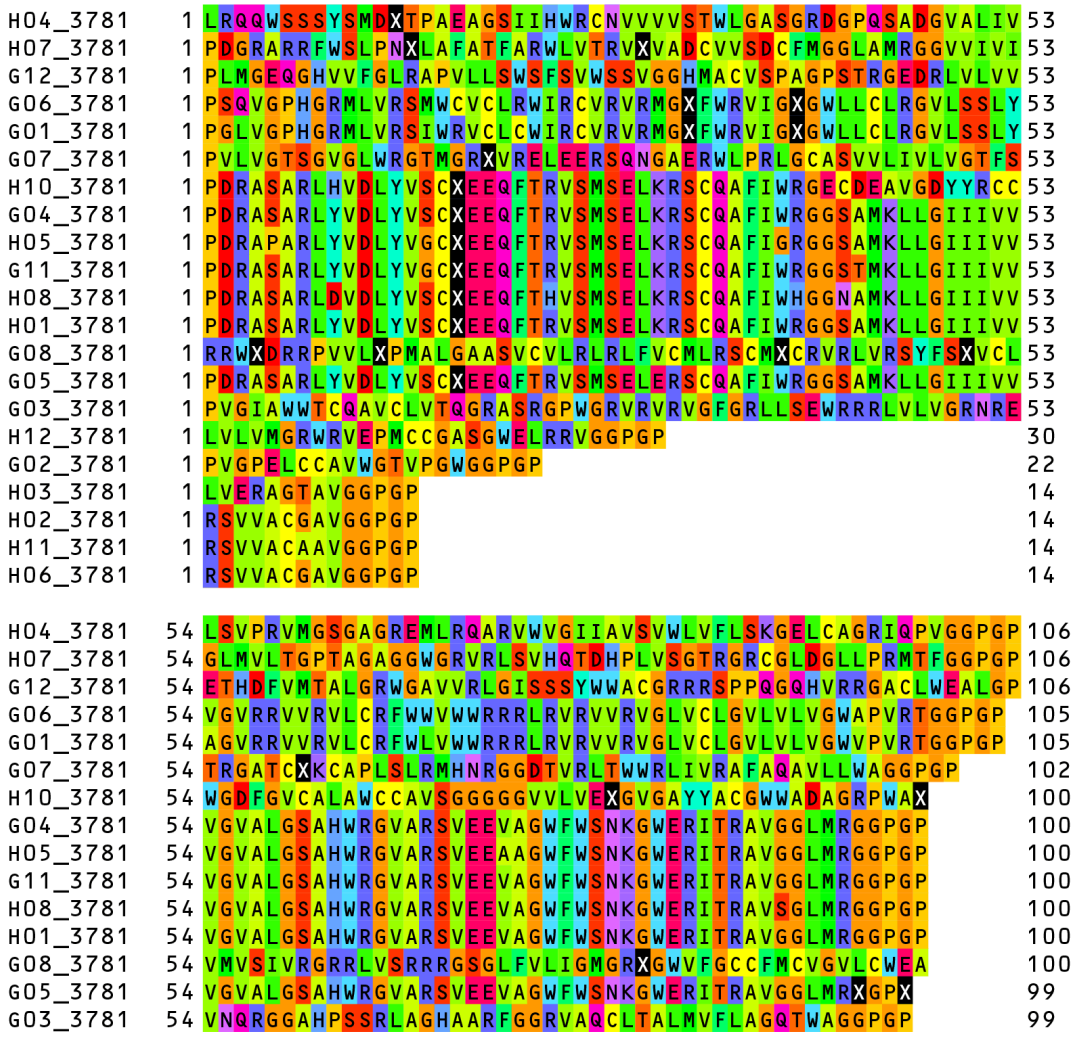


Figure 2.44: Sequencing results after round 5.
Most sequences contained stop codons, one dominant construct of 95 amino acids, containing the central hydrophobic patch.

2.5.9 Extensive screening by periplasmic fluorescence

The initial characterization of the output after round 5 showed a large proportion of constructs with stop codons. None of these constructs longer than 50 amino acids lead to measurable periplasmic localization of fluorescence by the fused reporter SF·GFP-SclpX. Therefore, a large-scale screening after selection round 5 was conducted.

The screening for periplasmic fluorescence was performed in DH5α using the TorA_ss–SF·GFP-SclpX format. Both outputs of round 5, of the selection with medium stringency at 50 µg/ml carbenicillin and with high stringency at 70 µg/ml, were cloned separately into the screening vector.

The output of round 4 was not re-cloned, as sorting in round 4 was performed in the TorA_ss–SF·GFP-SclpX format.

About 19% of 365 analyzed single colonies after round 4 showed significant fluorescence in the periplasm, all due to the fusion of a short peptide instead of a full-length insert. This shows that

short insert sequences get enriched in each round and the re-cloning of the library by gel extraction of the full-length band is pivotal to circumvent a massive selection bias towards short peptides.

More than 750 single colonies from round 5 were screened for periplasmic fluorescence by measuring the periplasmic fraction after cold-osmotic shock preparation. 28 constructs were chosen for sequencing according to their periplasmic fluorescence, which was in the range of 5% to 30% of the positive control. The fluorescence remaining in the spheroblasts was also recorded and used as an additional ranking criterion. Of the 28 sequenced constructs 9 contained stop codons. Two sequences were each found twice, resulting in 19 unique constructs.

2.5.10 Expression tests as SF·GFP fusion proteins

To test if any of the 19 constructs without stop codon from the screening for periplasmic fluorescence after selection round 5 (**Figure 2.45**: e01-e19) can be expressed as soluble protein, an expression vector containing a fusion protein of SF·GFP with a C-terminal 10 His-tag was constructed. As controls, the dominant constructs found in selection rounds 6α and 5, one unselected sequence of the random library, and lambda phage protein D were cloned into the expression vector.

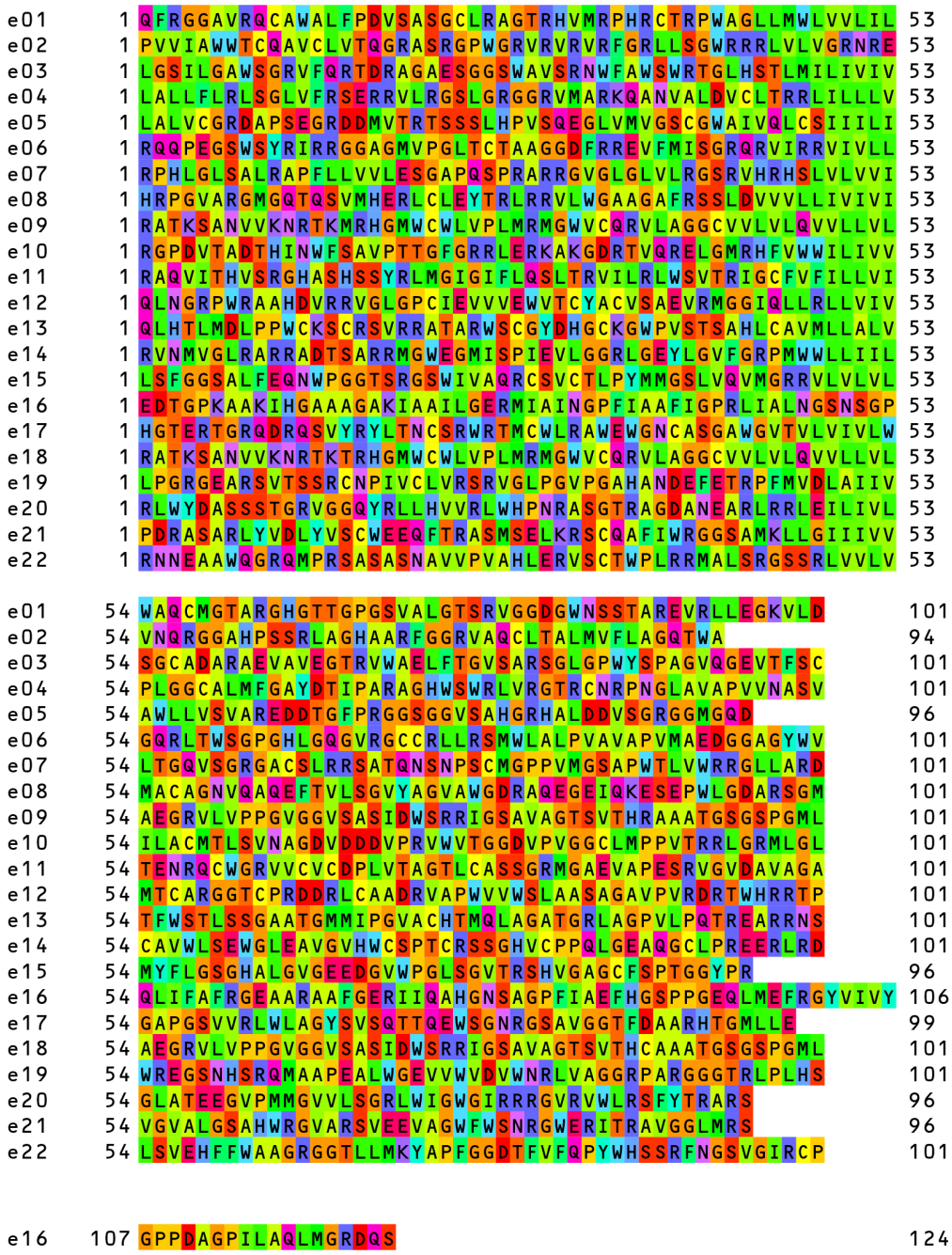


Figure 2.45: Alignment of the 22 constructs used for expression tests.
e09 and e18 originate from the same sequence and have 99 identical residues (of 101), e02 and e16 do not contain the designed hydrophobic patch, e20 is the dominant constructs from round 6α and e21 the dominant constructs from round 5 without stop codon, most constructs consist of the designed 101 amino acids.

Expression was performed in BL21 with 100 µM IPTG for 14 h at 23°C. The soluble fraction after fast prep with glass beads was loaded on SDS poly-acrylamide gels and in-gel fluorescence of SF·GFP was recorded.

All constructs, except for the positive control gpD, showed a fluorescent band at a height corresponding to SF·GFP only, or SF·GFP fused to short peptides. Construct 4 and 16 showed an additional weaker band that could correspond to full-length protein fused to SF·GFP (**Figure**

2.46). Loading of whole cells lysed in SDS loading buffer also showed fluorescent bands at the height of SF·GFP only, indicating that the POIs are not cleaved of by proteolysis during cell disruption.

The mass of full length fusion proteins would be around 40 kDa and their bands were expected noticeably above the 37 kDa marker band, yet the observed bands located close to the 25 kDa marker band, which would rather match bare SF·GFP with 28 kDa.

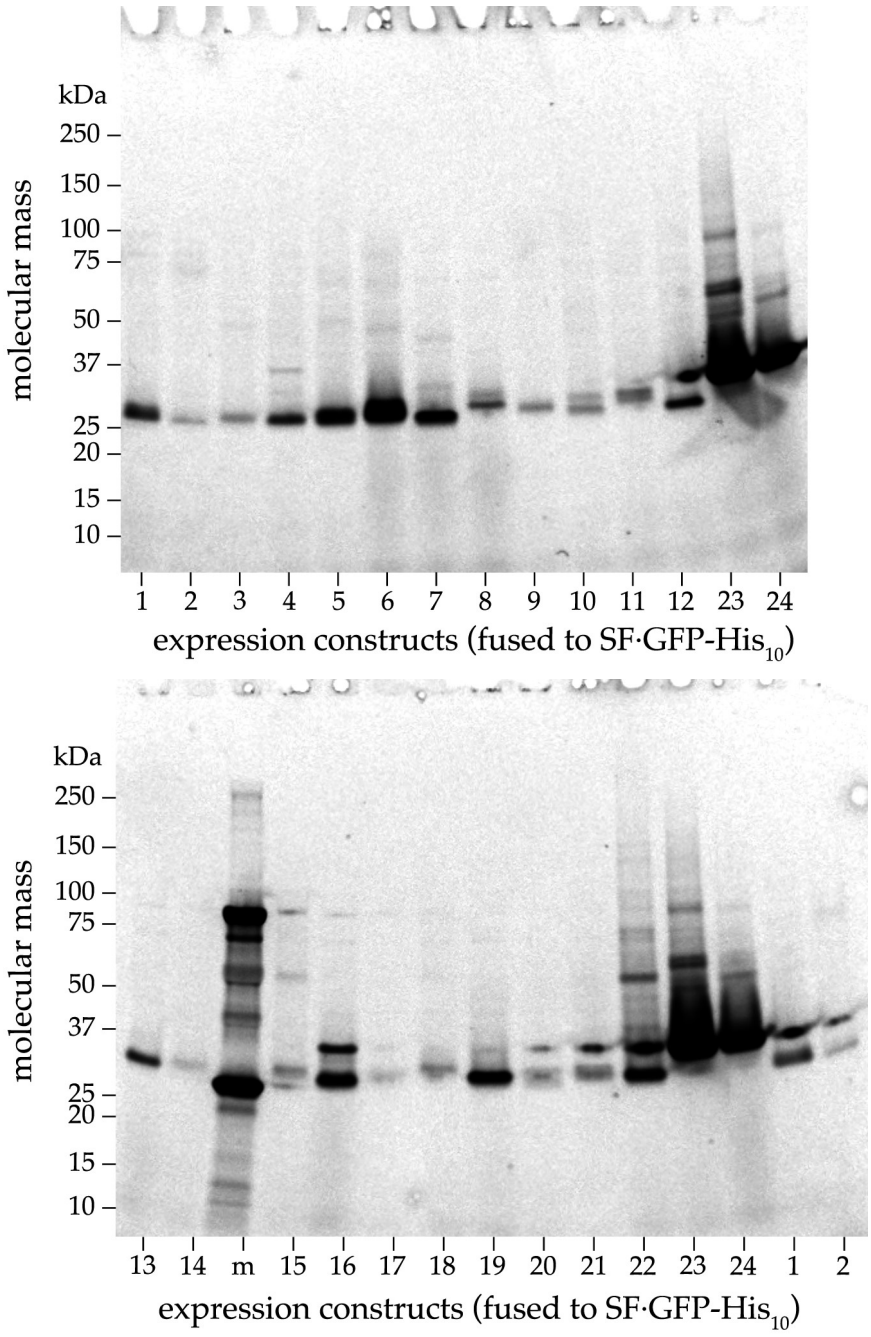


Figure 2.46: In-gel fluorescence scans of constructs expressed as SF·GFP fusion proteins. Construct 1-19 were identified in the screening for periplasmic fluorescence after round 5, no. 20 and 21 are the dominant constructs found in selection rounds 6α and 5, construct 22 is an unselected sequence without stop codon, 23 and 24 are gpD, lane "m" is the protein ladder "Precision Plus Dual Color Standard" from Bio-Rad Laboratories; the mass of the fused SF·GFP-His10 is 28.2 kDa, gpD-SF·GFP-His10 has a mass of 38.3 kDa.

The detection of SF·GFP only or SF·GFP fused to short peptides could be due to a cryptic translation start in the sequence of the POIs, close to the fused SF·GFP. Alternatively, the POIs could be rapidly degraded by cytoplasmic proteases during expression due to the high number of basic residues in their sequences, and possibly not being in a stably folded state. SignalP 3.0 did not recognize any of the constructs as potential signal sequence, which could have explained the periplasmic fluorescence as result of translocation via the SRP pathway, bypassing Tat-dependent translocation.

Further research is needed to elucidate the mechanism by which these proteins showed significant periplasmic fluorescence in the screening after round 5, which lead to their classification as promising candidates. Their Tat-dependent translocation should be confirmed in e.g. the Δ*tatC* strain and the length of exported proteins should be checked by e.g. immuno-blotting or in-gel fluorescence before examining these construct in more detail or trying to characterize the purified proteins without fusion partner.

3 Discussion

The search for truly novel proteins is as intricate as it is fascinating. In this thesis I tried to tackle the two most important aspects of this search: generating valuable libraries that cover unexplored sequence space and developing a potent selection system targeting stable protein folding.

The Tat pathway achieves the transmembrane passage of completely folded proteins and its folding quality control holds great promise for an in vivo selection system that may be able to address protein folding directly.

We measured the translocation efficacy of the Tat pathway in comparison to the other major translocation pathways in *E. coli* and, more importantly, in correlation to the folding properties of the Tat-targeted proteins. Two distinct reporter proteins were tested and the parameters for their Tat-dependent translocation were refined to identify the prerequisites, which have to be fulfilled to attain a potential selection system for folding.

To evaluate the full selection capability of our Tat-dependent reporter setups, we generated two libraries suitable for selections towards truly novel proteins. The binary patterned secondary structure library, which was previously constructed in our laboratory, was joined upstream and downstream of a central module encoding a hydrophobic patch of five amino acids to prevent the translocation of unfolded polypeptides via the Tat pathway. The other, fully random, library was created from scratch, where we were able to seamlessly assemble modules containing NNK codons and the module encoding the hydrophobic patch.

Finally, these libraries were used for selections with both established Tat-dependent reporter setups, giving rise to astonishing escape mechanisms, but also showing the promise as well as the limitations of these tools in the search for truly novel proteins.

3.1 *E. coli* translocation pathways and a putative selection system for folding

To measure the translocation efficacy of the Tat pathway in comparison to the other major translocation pathways in *E. coli*, we used the signal sequences phoA_ss targeting the SecB pathway and DsbA_ss targeting the SRP pathway for a primary analysis of two signal sequences targeting the Tat pore, TorA_ss and SufI_ss.

The main aim was to evaluate the potential of the Tat pathway for selections towards well-folded proteins. Therefore, the signal sequences phoA_ss and DsbA_ss were not used beyond initial investigations of the translocation of POI-reporter fusions via different *E. coli* export pathways, using Bla as reporter protein.

In a first attempt to establish a screening assay, the 96-well absorbance recordings of cell growth with the Bla setup in liquid culture looked practical. However, the strong dependence on shaking conditions and falsification of absorbance recordings caused by dead *E. coli* cells blocking the optical path lead to low reproducibility and made this assay unusable for our purpose. Export assays using the Bla setup in liquid culture were therefore replaced by assays on selective solid media plates.

In the Bla assay on solid media plates, DsbA_ss lead to efficient periplasmic export, via the SRP pathway, of almost all POI–Bla fusions even under stringent conditions with 100 µg/ml ampicillin. Only the aggregation-prone negative control AE564 showed a strongly reduced export, which resulted in the growth of colonies only in the second dilution step, whereas the fibril former with the worst solubility 3ex24 surprisingly lead to growth of colonies up to dilution step 7. All other constructs showed no limitation of SRP-dependent protein export, forming colonies up to the last dilution step 8 (**Figure 2.8b**).

The translocation via the SRP pathway occurs co-translationally, but the N-terminal region of a protein carrying a SRP signal sequence as DsbA_ss may already get translated by the ribosome until SRP binds to DsbA_ss and halts the translation to direct the complex to the Sec translocon.

3ex24 may need more than its N-terminal region for multimerization and the formation of fibrils, whereas the N-terminal region of AE564 may be sufficient to establish its strong oligomerization tendencies. This could prevent efficient SRP-dependent translocation of AE564.

Proteins with phoA_ss are routed to the Sec translocon post-translationally and the fully translated polypeptide is kept in an unfolded, extended conformation by binding of SecB. Constructs prone to oligomerization or aggregation, as 3ex24 or AE564, did not translocate to the periplasm efficiently, when targeted to the SecB pathway by phoA_ss. Compared to DsbA_ss, growth of colonies for these constructs was observed only in lower dilution steps and the reduced export was observed on low stringency plates as well. Notably, we did not observe a measurable reduction of export rates for other constructs with oligomerization tendencies or reduced solubility, as 2ex10 or 4ex24, when using phoA_ss.

The other mechanism preventing efficient SecB-dependent translocation of constructs with phoA_ss can be seen for very fast folding proteins as TrxA or the DARPin E3_5, where SecB probably cannot keep the polypeptide in an extended and translocation-competent state. This effect is stronger for the artificial protein E3_5 and was previously found to be the main reason hindering the incorporation of DARPins in SecB-dependent phage assembly [174].

In addition to the primary translocation analysis of POI-Bla fusions, phoA_ss and DsbA_ss were used to study possible export routes of fluorescent SF·GFP to the periplasm. Translocation of GFP resulting in periplasmic fluorescence had been reported only for Tat-dependent export. In our tests of targeting SF·GFP to the periplasm without additional fusion proteins we evaluated all four signal sequences. Surprisingly, the SRP-dependent DsbA_ss lead to the translocation of SF·GFP to the periplasm resulting in periplasmic fluorescence. It was later confirmed in an independent study that fluorescent SF·GFP can be exported to the periplasm using a modified MBP signal peptide, which is recognized by SRP and promotes cotranslational export [139].

The two Tat-dependent signal sequences TorA_ss and SufI_ss generally showed reduced periplasmic export rates for the Bla fusion proteins, compared to phoA_ss or DsbA_ss. The expression conditions used for characterization were comparatively mild not to oversaturate the Tat transport capacity, as the Tat pathway is limited in its throughput [108] compared to the Sec pathways, without overexpression of certain Tat-pore proteins [109,175,176]. In comparison to

TorA_ss, constructs with SufI_ss showed very low export rates for Bla fusion proteins, and no measurable fluorescence for SF·GFP, with and without additional proteins fused before SF·GFP. Lower Tat-dependent translocation rates of proteins with SufI_ss were reported independently [79], especially for GFP fusions [175,177,178].

We first used β-lactamase as reporter protein to estimate the potential of the Tat pathway to discriminate proteins based on their folding behavior. Export rates of POI–Bla fusion proteins targeted to the Tat pore by SufI_ss and more pronounced by TorA_ss did correlate to a certain degree with the folding characteristics of the POI. This allowed a phenotypic readout where a more stable folding leads to higher cell survival due to the increased periplasmic localization of the fused reporter protein β-lactamase, conferring resistance to β-lactam antibiotics.

The positive controls TrxA, gpD, and E3_5 showed the highest export rates, where the natural proteins TrxA and gpD were efficiently translocated with TorA_ss and the artificial, designed ankyrin repeat protein E3_5 showed a somewhat lower export rate, as seen for high stringency selective plates. The generally lower export rates with SufI_ss do not allow a fine-grained differentiation. A notable observation is that the DARPin E3_5 showed a comparatively good export rate, whereas the translocation efficiency of gpD was reduced with SufI_ss on low stringency selective plates.

A conclusive distinction of the folding characteristics of the four artificial armadillo repeat proteins Y2CA, Y2MA, Y4CA, and Y4MA based on their Tat-dependent translocation as Bla fusion proteins is not attainable. The previously untested short variant with two internal repeats based on the consensus design, Y2CA, showed a slightly better Tat-dependent translocation compared to Y2MA, which contains the mutated internal repeat modules. These mutated modules were reported to significantly improve the folding for armadillo repeat proteins with 3 or 4 internal repeats [134,135]. The well characterized improvement regarding stable folding of Y4MA compared to Y4CA was one of the reasons to include these proteins in the set for testing the Tat pathway. However, as Bla fusion proteins targeted for translocation via the Tat pathway, largely no significant differences could be observed. These artificial armadillo repeat proteins might be in general not well compatible with Tat-dependent translocation, at least not with the tested signal sequences.

Of the proteins from the secondary structure library, none showing a stable monomeric folding, the shortest variant AE73 stands out. It was exported unexpectedly well as Bla fusion protein with TorA_ss, showing that short unfolded polypeptides can bypass the quality control of the Tat pore, at least in the format of the employed Bla fusion construct.

Anti β-lactamase western blots of POI–Bla construct with TorA_ss and SufI_ss showed that most constructs can be expressed at comparatively high rate, with the exception of constructs with low solubility or unfavorable folding properties, especially the proteins from the SSL. No direct correlation of protein levels and export potential is visible, as e.g. SufI_ss–gpD–Bla expressed at high levels but showed lower export rates than SufI_ss–E3_5–Bla (**Figure 2.9** and **Figure 2.8**).

The expression levels of AE73–Bla were lower than for the armadillos, yet TorA_ss–AE73–Bla performed better in the export assays.

In addition to the Tat-dependent export of AE73–Bla, where a short unstructured polypeptide could bypass the folding quality check of the Tat system, the results of the first library selection using SSL2.1 fused to Bla (2.4.1) showed an even greater shortcoming of the Bla setup. β-lactamase is functional in the periplasm irrespective of the translocation pathway employed for export and the Tat pathway has the lowest throughput of the major translocation systems in *E. coli*. As a consequence, the probability of selecting signal sequences that target the Sec translocon may be much higher than finding truly novel, stably folded proteins by a Tat-dependent translocation in the Bla setup.

To prevent such a bypass of the desired selection pressure, a different type of reporter protein is needed, whose phenotypic readout depends on the translocation exclusively via the Tat pathway and the export via alternative routes leads to a loss of phenotype.

Translocation of GFP to the periplasm in a way that results in a properly folded and fluorescent protein had been reported only via the Tat pathway but not via the Sec translocon [109,112-114].

The discovery that SRP-dependent translocation of superfolder GFP leads to fluorescently active protein in the periplasm (2.1.7) has been surprising, as it can be observed for no other variant of GFP but only SF·GFP [139]. Thus, SF·GFP, but no other variant, can fold properly in the periplasm. The periplasmic fluorescence signal of SF·GFP translocated via the Tat pathway with TorA_ss was approximately twice as high (**Figure 2.10**, Periplasm) as the one from SRP-dependent translocation with DsbA_ss and SF·GFP showed the highest dynamic range for correlation of fluorescence with Tat-dependent export linked to protein folding, especially with the degradation tags necessary for elimination of cytoplasmic fluorescence. Therefore, we chose SF·GFP as reporter protein for Tat-dependent selections.

In initial tests, the periplasmic fluorescence readings for TorA_ss–POI–SF·GFP without degradation tag (**Figure 2.11**) showed a significantly higher fluorescence for Y4MA compared to Y4CA. Such strong differences were not measurable neither in GFP setups with degradation tag nor in the Bla setup for most of the tested conditions. The measurement of ssrA-tagged constructs in deletion strains for components of the degradation machinery (2.1.8) may provide a possible explanation. There, TorA_ss–Y4MA–SF·GFP-ssrA showed a measurably higher periplasmic localization in the Δ*clpX* strain (**Figure 2.13b**) compared to Y4CA, indicating that the artificial armadillo repeat proteins might need a considerable amount of time, prior to export, to fold correctly in the cytoplasm, where the rather quick operation of the ClpXP degradation system will remove all tagged proteins, which do not pass the folding quality check of the Tat pathway within a certain time frame. Yet, this does not fully explain the results obtained in the Bla setup, where such a strong discrimination of Y4MA and Y4CA is only visible at 30°C with low stringency plates having 25 µg/ml carbenicillin (**Figure 2.38**).

The unstructured polypeptide AE73, the shortest member of the test proteins from the secondary structure library, was exported surprisingly well via the Tat pathway when fused to β-lactamase

(**Figure 2.8**, **Figure 2.38**). However, AE73 could no longer bypass the folding quality control of the Tat pathway when fused to SF·GFP.

For selections based on flow cytometry sorting it is necessary to be able to use the fluorescence readout of whole cells as measure of Tat-dependent translocation of well folded proteins fused to e.g. SF·GFP. These SF·GFP fusion proteins are synthesized by ribosomes in the cytoplasm, where especially the domain of superfolder GFP adopts its native tertiary structure, including the formation of the fluorophore, comparatively fast whereas most proteins encoded by a library covering unexplored sequence space will presumably remain unfolded or take longer time to establish interactions that lead to the formation of secondary or tertiary structures. The results of the fluorescence localization for TorA_ss–SF·GFP (**Figure 2.10**) show that a considerable amount of GFP is located in the cytoplasm and will contribute to the fluorescence signal of whole cells significantly. To retain only the periplasmic fluorescence signal and be able to correlate whole-cell fluorescence with Tat-dependent translocation, cytoplasmic GFP has to be eliminated efficiently. However, proteins targeted for export to the periplasm via the Tat pore show an extended retention time in the cytoplasm prior to transport [107,108]. Consequently, the two systems, Tat-dependent translocation to the periplasm and cytoplasmic degradation of proteins that fail to be exported, need to be balanced carefully.

We chose the ClpXP machinery for proteolytic degradation of cytoplasmic proteins carrying a C-terminal ssrA peptide tag or a related sequence, based on a previous study [118].

In cells with a fully functional SspB/ClpXP degradation machinery, SF·GFP fusion proteins carrying a C-terminal ssrA tag were cleared very rapidly in the cytoplasm. Proteins in the format of TorA_ss–POI–SF·GFP-ssrA, targeted for Tat-dependent translocation, showed a weak fluorescence signal only for the positive control TrxA and no fluorescence for cells expressing other test proteins.

We sought to improve the dynamic range of the GFP system by increasing the phase before cytoplasmic degradation in order to give the Tat proofreading mechanism the needed time to export stably folded proteins. Expression of ssrA-tagged constructs in deletion strains missing one of the components of the ClpXP protease complex showed the strongest effects for Δ*clpP*, where the peptidase ClpP is not present, leading to a large increase in fluorescence, mainly located in the cytoplasm. With the ClpX unfoldase missing, Δ*clpX* showed less severe effects, where primarily the armadillo Y4MA and TrxA showed increased cytoplasmic fluorescence recordings. The two deletion strains Δ*clpP* and Δ*clpX* turned out not to be suitable for selections based on export via the Tat pathway, as the proteolytic degradation of tagged proteins in the cytoplasm becomes overly delayed, generating significant fluorescence signal, which is not located in the periplasm.

ClpXP-dependent degradation of ssrA-tagged proteins is enhanced by binding of SspB to its recognition sequence stretch in the C-terminal peptide tag [121,122]. In the strain Δ*sspB*, missing the adapter protein SspB, only the positive control TrxA showed a small fluorescence signal located in the cytoplasm (**Figure 2.13**), which may be due to expression levels that are slightly above the capacity of Tat-dependent translocation. Additionally the fluorescence levels recorded

in flow cytometry showed a better separation of positive and negative controls, especially at lower temperatures (**Figure 2.13**). Using the Δ*sspB* strain, removal of SspB was found to allow the necessary modulation of ClpXP-dependent degradation rate with respect to increased export of well folded proteins via the Tat pathway due to slower recognition and degradation of ssrA-tagged proteins in the cytoplasm, leading to higher fluorescence signals located in the periplasm.

All deletion strains were acquired from the Keio collection. Δ*uvrA* was used as ClpXP unrelated control, derived from the same parent as all Keio strains, BW25113, having been exposed to similar selection and growth procedures as Δ*sspB*, Δ*clpX*, and Δ*clpP*. Interestingly, the strain Δ*uvrA* showed a better separation of the constructs compared to the laboratory strain TOP10F' used for the initial characterization of SF·GFP-ssrA constructs or the stain DH5α used for characterization of the mutated tags and in library selections.

After identifying that removing the SspB:ssrA interaction in ClpXP-dependent degradation allows sufficient modulation for increased Tat-dependent translocation, we decided to use the recognition sequence in the C-terminal peptide tag to fine-tune the degradation rate. Thereby we avoid systemic perturbation of the cellular ClpXP degradation machinery, which the removal of a major component such as SspB could generate. This further allows the use of laboratory strains better suited for achieving high transformation rates needed for libraries with high diversity, as well as using suppressor strains for Amber stop codons, potentially helpful in selections using the random library based on NNK codons.

The C-terminal tags SclpX and Sprc were designed not to be recognized by the adapter protein SspB, while still targeting tagged proteins for ClpXP degradation. Analysis of proteins carrying these tags supported the use of a weakened degradation tag instead of the SspB deletion strain. The tags SclpX and Sprc show higher fluorescence for the positive control TrxA, both in DH5α and the deletion strains Δ*sspB* and Δ*prc*.

The positive control TrxA with one of the weakened degradation tags, e.g. TorA_ss–TrxA–SF·GFP-SclpX, showed similar or higher fluorescence when expressed in DH5α compared to expression of TorA_ss–TrxA–SF·GFP-ssrA, carrying the strong degradation tag, in Δ*sspB*. In both cases the fluorescence was detected only in the periplasm. Furthermore, there was little or no measurable increase in fluorescence when SclpX-tagged proteins were expressed in the strain lacking the adapter protein SspB. These are strong indications that the mutated sequence of SclpX and Sprc is no longer recognized by SspB and that the contribution of the SspB:ssrA interaction to ClpXP-dependent degradation can be modulated equally well or even better by modification of the degradation tag itself.

Expression of the aggregation-prone negative control AE564 with a weakened degradation tag at higher levels, 37°C and 200 µM IPTG, revealed a mechanism of generating false positive signals in the GFP setup with cytoplasmic degradation. If a protein, fused to GFP with a C-terminal degradation tag, is prone to aggregation, oligomerization or shielding of the degradation tag in a way that prevents cytoplasmic degradation but preserves the fluorescence of the fused GFP, and with these properties is unlikely to be exported via the Tat pathway, a significant fluorescence

signal will be generated, which is located predominantly in the cytoplasm and is in no way correlated with Tat-dependent translocation or stable, monomeric folding of this protein.

This mechanism of generating false positive signals in the GFP setup had been first observed for AE564 and Y4MA mainly with a weakened degradation tag at higher expression levels. However, the false positive constructs obtained from selection round 2α using the random library (2.5.3) showed that proteins could be selected, where this mechanism works with the strong degradation tag ssrA at milder expression conditions and even with the wild-type like S65T·GFP. In all cases the fluorescence signal was located predominantly in the cytoplasm. Later selection rounds in the GFP setup on the random library were most likely affected by this mechanism as well.

Regarding their potential for selections based on Tat-dependent translocation both reporter proteins showed certain strengths and limitations.

The Bla setup may be of use for selections against oligomerization and aggregation and can partially complement GFP-based selections, as in the case of the false positive constructs from round 2α (2.5.3), where some constructs showed reduced export rates as Bla fusions. Yet, the Bla setup allows Tat-dependent export of unstructured polypeptides like AE73, even if they contain a hydrophobic patch as in the case of A11 (2.4.5, **Figure 2.38**). In general, Tat-dependent translocation seems to be more permissive for Bla fusion proteins, indicating a less strict folding quality control for this setup. Additionally, functional export of β-lactamase to the periplasm is not restricted to the Tat pathway and a polypeptide that is able to direct the fused Bla to be exported e.g. via the Sec translocon will easily dominate selections based on the Bla setup (2.4.2).

The GFP setup, carefully balanced for Tat-dependent translocation to the periplasm versus removal of remaining cytoplasmic fluorescence, shows a large dynamic range and may be better suited for selections concerning the steady progress of protein folding characteristics. In the GFP setup, the short unstructured polypeptide AE73 does not display significant export to the periplasm. Tat-dependent translocation and its protein folding quality check of the fusion proteins seem to be more stringent in the GFP setup than for Bla fusions.

Export of fluorescent GFP to the periplasm was thought to be Tat-exclusive [109,112-114]. Despite of the surprising discovery that SRP-dependent translocation of SF·GFP leads to fluorescent protein in the periplasm (2.1.7, [139]), we weighed the greatly improved fluorescence properties of superfolder GFP [117] higher than a potential risk of bypassing Tat-dependent translocation to the periplasm via generation of SRP signal sequences. Consequently, Tat-dependent translocation of promising candidates identified in the GFP setup should always be tested in e.g. *tatC* deletion strains.

The highest risk of generating false positives in the GFP setup has been identified in constructs that do not show significant Tat-dependent export but accumulate in the cytoplasm with fluorescent GFP and inaccessible degradation tag. Counter-selections against these constructs based on the Bla setup worked only partially and, although SF·GFP-Bla fusions coupled with selections for resistance to β-lactam antibiotics showed further improvements, not all variants could be eliminated (2.5.3). The lower export rates of some constructs as SF·GFP-Bla fusions

may be explained by the overall increase in molecular mass or by the influence of the immediate protein vicinity on oligomerization tendency, as these constructs were selected as SF·GFP fusions (with TorA_ss and SclpX tag, see 2.5.2).

3.2 Generating libraries useful for selections towards truly novel proteins

Tat-dependent translocation, using suitable reporters, might be a potent tool for the identification of stably folding proteins. To use it for selections towards truly novel proteins, libraries of proteins are needed that show little to no homology to natural occurring protein domains.

The amount of DNA or genotype-phenotype linkages that can feasibly be handled in experiments limits the number of variants that can be screened at any one time. For in vitro selections the upper limit is around 10^{14} variants [49,150], whereas the limit for in vivo selections is a few orders of magnitude lower at 10^{10}, mainly restrained by transformation rates of the DNA into the organisms.

For a typical size of a protein domain [46-48] consisting of 100 amino acid residues, the theoretical diversity of a library will quickly exceed the number of variants that can be handled experimentally, even if not all positions are variable or the number of allowed residues per variable position is restricted. The intuitive reaction is often to reduce the theoretical diversity. This approach may be sensible for libraries where variants of a stably folding protein are generated either by introducing random mutations (e.g. by epPCR) or by generating designed variability in chosen positions (e.g. by wobble codons or trinucleotide mixes); whether reducing the theoretical diversity is helpful for libraries with a large number of randomized positions remains questionable (see 2.3).

In the case of a known scaffold it is often possible to anticipate the amount of tolerated mutations regarding the conservation of a stable folding by comparison of homologous sequences. The information from a solved structure further allows to formulate a set of unfavorable residues which to exclude in the library design, based on the context of the known local contacts. This approach is mostly based on information obtained from solved structures and works reliably only when the number of involved residues is not larger than a few dozen.

For a protein of about 100 amino acids there is no obvious way to formulate a set of rules that will lead to a high propensity of finding a truly novel, stably folded protein. The growing number of solved protein structures does provide an increasingly robust set of (super-secondary) structural elements, including their crucial interaction aspects, which can then be transferred to other structural contexts to assemble stably folding proteins. Nevertheless, the underlying principles governing protein folding are still far from being well understood.

Today, the most promising approach to obtain a novel protein that has a high propensity of stable folding would be to use these elements derived from solved structures. This brings about a high local homology to known protein folds and is unlikely to result in a truly novel protein fold, rather than domain-swapped variations or other known folds; which is not surprising if we assume that the evolution of natural proteins has been driven by similar mechanisms.

One impressive example, where these structural elements have been used to successfully generate a truly novel protein, is the computationally optimized designed structure TOP7 [43]. So far it has remained the only published example of this procedure.

The question whether this stably folding protein, TOP7, could be selected from a combinatorial library consisting of a random shuffling of its secondary or super-secondary structural elements and if such a library would contain other stably folding proteins, remains open.

A technique to reduce the theoretical diversity of a library and at the same time have the highly promising attribute of forming secondary structure elements, was introduced with the binary patterning principle of polar and non-polar residues [51,52]. At first this sounded like a set of rules that would lead to libraries with high propensities of finding truly novel proteins.

However, reports indicated that proteins with binary patterned beta strands show high oligomerization and aggregation tendencies [53,54,56,57], while alpha helical modules based on binary patterning worked better and in some cases showed formation of the respective secondary structures [51,55].

In addition to the basic shortcomings in the binary patterning design, our implementation, the secondary structure library, shows one major practical drawback that makes it almost unusable in applications that include any PCR steps. The high homology within variants of one module coupled with the randomized, shuffled arrangement of the modules leads to a continuous shortening of the whole library due to hybridization of similar modules from different DNA strands and generation of undesired shortened side-products, which become amplified more efficiently with increasing PCR cycles.

To create an alternative to the SSL without such practical and conceptual limitations, we devised the construction of a fully random library. This random library should further not be biased in its covered sequence space, especially as the bias of binary patterning showed only partial correlation of concept and experimental results.

It is difficult to estimate if the number that can be screened in vivo, of 10^9 variants in our case, is sufficient to have a reasonable probability of finding (partially) folded proteins in a fully random library. In the one published example of a successful selection of an ATP binder using a random library, Keefe et al. were able to screen more than 10^{12} variants [44].

Maybe an in vitro selection step will have to precede any in vivo selections, in order to have a higher quality library for transformations, consisting e.g. of the top 1% soluble, in-frame variants from an in vitro selection on a library with 10^{12} variants.

The iterations in the construction of the fully random library emphasize that an initial design should not be too complex and should not contain a large number of unknown parameters. Instead, robust and well established methods should be chosen whenever possible, and a stepwise implementation of the most important features will reveal further bottlenecks. For how these problems were eventually solved, see 2.3.

The addition of a hydrophobic patch to any library used in selections involving Tat-dependent translocation can help to prevent the export of unstructured polypeptides [106] and further function as nucleation center in the formation of a solvent shielded protein core [153].

The optimal positioning of the hydrophobic patch is difficult to predict. We incorporated it in the center of the libraries used for selections, but the construction scheme of the random library would allow any positioning of the hydrophobic patch in a completed library. This could as well be applied in a de-novo construction of a secondary structure library, whereas a reassembly of two SSL2.1 entities with a hydrophobic patch module would restrict the available positions to the center or the termini.

3.3 Selections towards truly novel proteins

Selections for Tat-dependent translocation using the SSL and full random library revealed some common, uncommon and utterly unexpected escape mechanisms.

Short peptides fused between the signal sequence and the reporter are not rejected by the folding quality check at the Tat pore, as the fused reporter protein is stably folded and the peptide is just a minor addition to its shape or mass. Switching to a much smaller reporter, like a peptide tag, would be an option, as long as its functional export to the periplasm is Tat exclusive and the probability that a comparable tag is contained in the library is smaller than the occurrence of a stably folded protein.

Re-cloning of full length inserts after each selection round is crucial to profoundly minimize the percentage of short peptide inserts. Even if there is only a minor percentage of short constructs, these short sequences will be efficiently enriched in each selection round and PCR reaction (about one fifth of the constructs after round 4 were short peptides, see 2.5.9).

Similarly, double-transformants (as A11, see 2.4.5) and mutations in the reporter sequence (as for clone 4.24, see 2.4.3) or other parts of the vector backbone can in most cases be eliminated as potential cause of false positives by re-cloning.

Equally, the different reporters come with limitations as well (discussed in detail before, see 3.1). The Bla setup showed a less stringent folding quality check for Tat-targeted translocation, as seen for AE73 and more so for A11, containing a hydrophobic patch (2.4.5). And the unexpected escape mechanism of the S7 insert (2.4.2) demonstrated the importance of coupling the translocation of a functional reporter as closely as possible to the Tat pathway and exclude other routes without folding quality check. While false positives such as S7 would be efficiently eliminated in the GFP setup, the establishment and fine-tuning of cytoplasmic degradation versus Tat-dependent translocation (2.1.8, 2.1.9) already indicated possible weak spots of the GFP setup. Constructs with significant cytoplasmic fluorescence in the GFP setup, which do not get exported but escape degradation by ClpXP despite carrying a degradation tag can confound the selection. The selection process of sorting cells in flow cytometry can merely measure whole-cell fluorescence, such that any signal that does not originate in the periplasm due to successful translocation via the Tat pathway, including its folding quality check, will lead to false positives.

The Bla setup does not fully complement the GFP setup in regard to counter-selections of false positives that evade cytoplasmic degradation as GFP fusions. This has been initially revealed with constructs from the stringent selection round 2α (2.5.3). And it is not utterly surprising that this type of escape mechanism could not be fully averted by lowering the selection stringency or by introducing SF·GFP-Bla as a reporter slightly better suited for counter-selections based on antibiotic resistance. The dominant constructs after round 6α and round 5 do not show significant periplasmic fluorescence, as they probably use similar escape mechanisms, where they evade cytoplasmic degradation despite carrying a specific degradation tag as GFP fusion proteins, yet they did not get eliminated in the Bla-based selection rounds.

To dominate selections using both the GFP and the Bla setup a protein could either adopt a stably folded structure or have the two following properties: The ability to evade cytoplasmic degradation in the GFP setup as well as getting sufficiently translocated to the periplasm in the Bla setup, maybe not even via the Tat pathway, to confer resistance to the applied antibiotic selection pressure.

Generally, the probability of finding construct that have both properties necessary to bypass the intended selection route is almost certainly greater that finding truly novel, well folded proteins.

3.4 Future perspectives

The constructs discovered in the screening after round 5, which, for the first time, did show significant periplasmic fluorescence certainly deserve further investigation. In silico analysis of their sequences by SignalP did not indicate that alternative signal sequences targeting the SRP pathway are encoded, which could also lead to fluorescent GFP in periplasm and thereby bypass the Tat folding quality check.

Yet, the fact that they could not be detected as SF·GFP fusion proteins in PAGE might be a sign of little or partial stable folding which would make them accessible for proteolytic cleavage in the cytoplasm and explain why only bands corresponding approximately to the size of bare SF·GFP were observed. It remains to be investigated whether such proteolysis may have as well occurred in the screening setup, leading to the translocation of truncated SF·GFP fusions.

Further experiments, e.g. for Tat-dependent translocation in the Δ*tatC* strain, or expression tests with other fusion proteins could help to examine these construct in more detail and maybe even characterize the proteins in purified form.

To obtain a useful implementation of the secondary structure library a whole reconstruction may be necessary, beginning from module level with a reduced intra-module homology. There is a very small chance that a revised set of flanking primer sequences might be found and established, which could possibly prevent the continuous shortening of the library in PCR reaction due to the hybridization of very similar variants of one module. With such a primer set, only a fresh assembly of an SSL3 type library would be needed, including the incorporation of the hydrophobic patch module in-between two SSL2.1 entities to block Tat-dependent translocation of unfolded polypeptides.

The random library shows little to no undesired side-products as long a certain precautionary measures are implemented.

Even if superior reporters can be found, a selection system based on Tat-dependent translocation probably needs to be complemented by a selection method not based on translocation, which is able to eliminate the false positive constructs observed here.

Such a selection system could conceivably recognize the folding state of proteins and would be a great tool, not only in the search for novel well-folded proteins. Together with a suitable library, a selection system for folding would be one of the most promising approaches today for finding truly novel proteins and thereby help to further unravel the principles that govern stable protein folding.

Looking at the evolution of protein folds, with few superfolds and many unifolds, and the examples of truly novel proteins reported, it is not unlikely that many more stable protein folds do exist.

4 Materials and methods

4.1 Materials

4.1.1 Oligonucleotides and DNA modifying enzymes

Oligonucleotides were purchased from Microsynth (Switzerland), DNA-modifying enzymes such as restriction enzymes, ligases, or DNA polymerases were obtained from New England Biolabs (USA) or Thermo Scientific (USA).

4.1.2 Bacterial strains

The *E. coli* deletion strains Δ*clpP*, Δ*clpX*, Δ*prc*, Δ*sspB*, Δ*tatC*, and Δ*uvrA* were obtained from the Keio collection [140], which is derived from the strain BW25113 (Sex: F- ; Chromosomal Markers: *Δ(araD-araB)567*, *ΔlacZ4787*(::rrnB-3), *λ⁻*, *rph-1*, *Δ(rhaD-rhaB)568*, *hsdR514*) and is listed with number 7636 in the CGSC (*E. coli* Genetic Stock Center at Yale University, USA).

Table 4.1: Strains from the Keio collection used in this work.

Short ID	Strain	CGSC#	Additional chromosomal markers
Δ*sspB*	JW3197-1	10426	Δ*sspB756*::kan
Δ*clpX*	JW0428-1	8591	Δ*clpX724*::kan
Δ*clpP*	JW0427-1	8590	Δ*clpP723*::kan
Δ*uvrA*	JW4019-2	10889	Δ*uvrA753*::kan
Δ*prc*	JW1819-1	9520	Δ*prc-755*::kan
Δ*tatC*	JW3815-1	10761	Δ*tatC781*::kan

Short ID is the naming used in this thesis, Strain gives the *Escherichia coli* strain name, CGSC# is the internal number in The Coli Genetic Stock Center (Yale University). Furthermore, the chromosomal markers additional to the ones of the parent strain BW25113 are listed.

Other strains were obtained from Life Technologies (Thermo Fisher Scientific) or New England Biolabs.

DH5α: CGSC# 12384; Sex: F-; Chromosomal Markers: *Δ(argF-lac)169*, φ80d*lacZ58*(M15), *ΔphoA8*, *glnX44*(AS), *λ⁻*, *deoR481*, *rfbC1*, *gyrA96*(NalR), *recA1*, *endA1*, *thiE1*, *hsdR17*. DH5α has a strong amber stop codon (UAG) suppressor in *glnX44*(AS) also known as *supE*, which leads to the production of the suppressor tRNA Glutamine tRNA2 that recognizes amber stop codons in competition with translation termination factors. Its efficiency of suppression by incorporation of glutamine is thus usually less than 100%.

4.1.3 Plasmids

Plasmids for characterization and selections were constructed from the template pDST22 [124], which is a derivative of the vector pMorph7 [125]. See **Figure 6.2** for a vector map of the main constructs used in this thesis.

4.2 Methods

Biochemical and molecular biology methods were performed according to standard protocols [179] and suppliers' manuals, with only few adjustments.

Ligations were incubated in a thermo-cycler for 1-2 h, running a endless loop of 12°C for 30 s and 30°C for 30 s.

Error-prone PCRs were performed with addition of 0.1 mM to 0.6 mM $MnCl_2$ to standard buffers using Taq DNA polymerase without proofreading activity.

Alkaline agarose gel electrophoresis was initially performed using alkaline gels, as described [179]. In the case of the SSL it turned out to be sufficient to use standard TBE or TAE gels and running buffers, and just add 1× to 2× alkaline gel-loading buffer before loading the DNA (add NaOH to standard DNA loading buffer in a final concentration of 50-100 mM).

Transformations of libraries were done by electroporation using 2 mM cuvettes (Eurogentec, Belgium) and ~70 µl of competent cells premixed with column-purified ligation product. Electroporation parameters were set according to the machine's manual. To obtain high numbers of transformed cells, up to twelve separate electroporations were pooled in 50 ml 2YT with 0.5% glucose and 25 µg/ml Cm for recovery.

4.2.1 Fast preparation of soluble protein fraction for in-gel GFP fluorescence

Fast preparation of soluble protein fraction (fast prep) was performed using glass beads and ultracentrifugation. For each construct, similar amounts of cells were sedimented corresponding to 1 µg total protein, where 1 µg of total protein is contained in roughly 3.8 ml of OD_{600} = 1. After discarding the supernatant, 400 µl buffer (50 mM Tris/KPi pH ~7.5 + 1 mM $MgSO_4$ + 150 mM NaCl + 10% glycerol + 1 mM PMSF + DnaseI) was added to the sedimented cells. Further, one “spoon” of glass beads (≤ 106 µm diameter) was added to each tube before performing the mechanical cell homogenization using the FastPrep-24 (MP Biomedicals, USA) with the settings: 6 M/s for 20 s. After an incubation for 5 min on ice, another round of fast prep with 6 M/s for 20 s followed. 140 µl of supernatant above the glass beads were transferred to small ultra-centrifugation tubes (due to volume of the beads, max. 200 µl supernatant can be recovered) and of these 140 µl, 40 µl were saved in separate tubes as whole-cell expression samples. The 100 µl remaining in the small ultra-centrifugation tubes were centrifuged in a Optima MAX-XP Ultracentrifuge (Beckman Coulter, USA) using the TLA-100 rotor for 10 min at 75'000 rpm ~250'000 rcf. 80 µl of the supernatant were transferred for analysis of the soluble protein fraction. Using 10 µl of the soluble protein fraction from fast prep, in-gel fluorescence of SF·GFP fusion proteins was recoded in the Fujifilm Image Reader LAS-3000 with blue light (460 nm EPI) for excitation and a Y515AttoPhos filter.

4.2.2 Bla solid-plate assay, droplet dilution series

Cells carrying constructs in the Bla setup were grown overnight at 33°C in LB medium containing chloramphenicol (25 µg/ml Cm). The overnight cultures, which had grown to the stationary phase of *E. coli* cell proliferation, were used to inoculate LB medium containing 25 µg/ml Cm and 0.5%

glucose. These freshly inoculated cultures were grown at 30°C to the exponential growth phase and at OD_{600} ~ 0.5 screening of cells on solid plates was performed by spotting 5 µl of each dilution step directly onto LB agar plates (25 µg/ml Cm, IPTG and β-lactam antibiotics according to intended stringency), similar to a previously published assay [107].

4.2.3 Bla selections in liquid medium using 5 µm filtration

After transformations by electroporation, the cells were recovered for 1.5 h at 33°C in 2YT with 0.5% glucose. Depending on the intended selection stringency, IPTG and β-lactam antibiotics were added (e.g. for round 1 of the MOAL selections, 100 µM IPTG and 20 µg/ml carbenicillin) and the cells were grown for another 2.5 to 4.5 h. Only transformed cells expressing Bla fusion constructs that can be translocated to the periplasm, namely via the Tat pathway, were able to undergo cell division and maintain a size that can pass through a 5 µm filter. This subpopulation, which represented a minor fraction of the total biomass, was mechanically separated from the bulk cell mass by passing the liquid culture through a 5 µm filter (Millipore). Cells unable to divide under selective conditions kept growing to long filaments and were retained in the filter. The cells that were able to pass the filter were sedimented by centrifugation at 5'000 rcf for 8 min and resuspended in fresh 2YT medium with 0.5% glucose and grown to mid or end of exponential growth phase. These cells were stored or used directly for further selections.

For storage, cells were centrifuged at 5'000 rcf for 8 min and resuspended in 1× Hogness modified freezing medium (HMFM): 36 mM K_2HPO_4, 13.2 mM KH_2PO_4, 0.4 mM $MgSO_4$, 1.7 mM Na_3-citrate, 6.8 mM $(NH_4)_2SO_4$, 12% (v/v) glycerol. Cells resuspended in 1×HMFM were snap-frozen in liquid N_2 and stored at -80°C.

4.2.4 Periplasmic extraction by cold-osmotic shock

Cold-osmotic shock was found to allow a fast and efficient preparation of the periplasmic fraction from *E. coli* cells. The following procedure describes the liberation of the periplasmic fraction in 96-well microtitration plates, which were used for large-scale screening of constructs (see 2.5.9). It can easily be adapted to larger volumes.

A volume of 800 µl of cells was sedimented using centrifugation with at least 4'000 rcf for 12 min. The supernatant was removed by pipetting (Mettler Toledo's Rainin Liquidator 96 Manual Pipetting System) and the cells were resuspended in 200 µl (1/4 starting volume) of 30 mM Tris·Cl, 20% (w/v) sucrose, pH 8 + 1 mM EDTA. After incubation at room temperature for 5-10 min with shaking, the cells were sedimented by centrifugation with at least 4'000 rcf for 12 min. The supernatant was removed and from this point on the microtitration plates were kept at ~4°C. The sedimented cells were resuspended in 200 µl (1/4 starting volume) of ice-cold 5 mM $MgSO_4$ by pipetting and incubated at ~4°C for 10 min, preferentially with shaking.

In a centrifugation step at ~4°C with at least 4'000 rcf for 15 min, the spheroblasts were sedimented. By pipetting, ~160 µl of the supernatant containing the cells' periplasmic fraction after the cold-osmotic shock were quickly transferred to supplementary 96-well microtitration plates containing 32 µl (1/5 of working volume) of 120 mM Tris·Cl at pH 7.4 per well. The sedimented spheroblasts were then resuspended in 1×PBS.

In the standard osmotic shock, the cells are first suspended in a concentrated solution of sucrose, which is supplemented with EDTA. After centrifugation, the cells are resuspended in cold water. The liberation of periplasmic proteins thus occurs. In this procedure, the sucrose solution of high osmotic strength causes the cells to shrink; EDTA functions to release lipopolysaccharide (LPS) from the bacterial cell envelope, thus increases the permeability of the outer membrane; finally, cold water causes a rapid increase in cell size, and results in the release of periplasmic proteins. Among the steps involved in osmotic shock, the increase of permeability is the key point if one wishes to improve the efficiency of periplasmic release.

The cell envelope of gram-negative bacteria contains three structures: a cytoplasmic membrane containing polar lipids and proteins; a peptidoglycan layer; and an outer membrane that contains 15–40 wt.% LPS, in addition to lipid and protein. The permeability barrier of the outer membrane has been shown to be related to the presence of LPS. LPS is stabilized by divalent cations, and the damage to the permeability barrier by EDTA treatment is due to chelation of divalent cations combined with a stripping-off of the LPS layer [180,181].

4.2.5 Flow cytometry screening and sorting

Fluorescence levels of constructs in the GFP setup depend on many parameters, e.g. IPTG concentration, temperature, expression time, bacterial strain, and type of degradation tag. Therefore, positive and negative controls should always be expressed in parallel, to facilitate accurate evaluation of recorded fluorescence levels for the conditions used, especially when handling libraries. As suitable positive controls TrxA and gpD have been used, and AE73 or AE564 as negative controls.

Before flow cytometry, cells were sedimented by centrifugation and resuspended in 1×PBS to an OD_{600} of ~1 and then diluted in 1×PBS to a concentration appropriate for the respective flow cytometer. From the time point of cell sedimentation, cells and buffers were kept on ice or 4°C. Resuspended cells were filtered according to the instructions for the respective flow cytometer.

The electronic amplification of the fluorescence signal, i.e. the voltage value in the parameters of the cytometer settings, was adjusted such that the maximum fluorescence of the positive control would not reach the upper detection limit but preferentially locate at about one to two orders of magnitude below that border. Additionally, the median of the fluorescence of the negative control should be positioned sufficiently above the lower detection limit.

4.2.6 Fluorescence measurements in 96-well plates

Fluorescence readings in 96-well plates were performed in a Infinite M1000 PRO microplate reader (Tecan, Switzerland) with an excitation wavelength of 470 nm. Emission was recorded from 498 nm to 554 nm. The bandwidth was set to 7 nm, for both excitation and emission. The folded positive controls of the SF·GFP setup had their emission signal maximum at 512 nm.

5 References

1. Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181: 223–230.
2. Levinthal C (1968) Are There Pathways For Protein Folding? Extrait du Journal de Chimie Physique 65: 44–45.
3. Levinthal C (1969) How to Fold Graciously. In: Debrunnder J, Munck E, editors. University of Illinois Press. pp. 22–24.
4. Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. Annu Rev Biophys 37: 289–316.
5. Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. Nat Struct Biol 4: 10–19.
6. Bryngelson JD, Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. Proc Natl Acad Sci USA 84: 7524–7528.
7. Ozkan SB, Wu GA, Chodera JD, Dill KA (2007) Protein folding by zipping and assembly. Proc Natl Acad Sci USA 104: 11987–11992.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242.
9. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10: 980–980.
10. Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261: 552–558.
11. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
12. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH – a hierarchic classification of protein domain structures. Structure 5: 1093–1109.
13. Kolodny R, Pereyaslavets L, Samson AO, Levitt M (2013) On the universe of protein folds. Annu Rev Biophys 42: 559–582.
14. Day R, Beck DAC, Armen RS, Daggett V (2003) A consensus view of fold space: Combining SCOP, CATH, and the Dali Domain Dictionary. Protein Sci 12: 2150–2160.
15. Hadley C, Jones DT (1999) A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure 7: 1099–1112.
16. Sippl MJ (2009) Fold space unlimited. Curr Opin Struct Biol 19: 312–320.
17. Chothia C (1992) Proteins. One thousand families for the molecular biologist. Nature 357: 543–544.
18. Govindarajan S, Recabarren R, Goldstein RA (1999) Estimating the total number of protein folds. Proteins: Struct Funct Bioinform 35: 408–414.
19. Orengo CA, Jones DT, Thornton JM (1994) Protein superfamilies and domain superfolds. Nature 372: 631–634.
20. Coulson AFW, Moult J (2002) A unifold, mesofold, and superfold model of protein fold use. Proteins: Struct Funct Bioinform 46: 61–71.
21. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5: 823–826.
22. Sander C, Schneider R (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins: Struct Funct Bioinform 9: 56–68.
23. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12: 85–94.
24. Goldstein RA (2008) The structure of protein evolution and the evolution of protein structure. Curr Opin Struct Biol 18: 170–177.
25. Kosloff M, Kolodny R (2008) Sequence-similar, structure-dissimilar protein pairs in the PDB. Proteins: Struct Funct Bioinform 71: 891–902.
26. Murzin AG (2008) Biochemistry. Metamorphic proteins. Science 320: 1725–1726.
27. Andreeva A, Murzin AG (2010) Structural classification of proteins and structural genomics: new insights into protein folding and evolution. Acta Crystallogr Sect F Struct Biol Cryst Commun 66: 1190–1197.
28. Alexander PA, He Y, Chen Y, Orban J, Bryan PN (2009) A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci USA 106: 21149–21154.
29. Hansen N, Allison JR, Hodel FH, van Gunsteren WF (2013) Relative free enthalpies for point mutations in two proteins with highly similar sequences but different folds. Biochemistry 52: 4962–4970.
30. He Y, Chen Y, Alexander PA, Bryan PN, Orban J (2012) Mutational tipping points for switching protein folds and functions. Structure 20: 283–291.
31. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the Protein Repertoire. Science 300: 1701–1703.
32. Choi I-G, Kim S-H (2006) Evolution of protein structural classes and protein sequence families. Proc Natl Acad Sci USA 103: 14056–14061.
33. Söding J, Lupas AN (2003) More than the sum of their parts: On the evolution of proteins from peptides. BioEssays 25: 837–846.
34. Orengo CA, Thornton JM (2005) Protein families and their evolution-a structural perspective. Annu Rev Biochem 74: 867–900.

35. Shakhnovich BE, Dokholyan NV, DeLisi C, Shakhnovich EI (2003) Functional Fingerprints of Folds: Evidence for Correlated Structure–Function Evolution. J Mol Biol 326: 1–9.
36. Salem GM, Hutchinson EG, Orengo CA, Thornton JM (1999) Correlation of observed fold frequency with the occurrence of local structural motifs. J Mol Biol 287: 969–981.
37. Cossio P, Trovato A, Pietrucci F, Seno F, Maritan A, Laio A (2010) Exploring the universe of protein structures beyond the Protein Data Bank. PLoS Comput Biol 6: e1000957.
38. Taylor WR, Chelliah V, Hollup SM, Macdonald JT, Jonassen I (2009) Probing the "dark matter" of protein fold space. Structure 17: 1244–1252.
39. Schwede T (2013) Protein Modeling: What Happened to the "Protein Structure Gap?" Structure 21: 1531–1540.
40. Bowers PM, Strauss CE, Baker D (2000) De novo protein structure determination using sparse NMR data. J Biomol NMR 18: 311–318.
41. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 97: 10383–10388.
42. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383: 66–93.
43. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. Science 302: 1364–1368.
44. Keefe AD, Szostak JW (2001) Functional proteins from a random-sequence library. Nature 410: 715–718.
45. Surdo Lo P, Walsh MA, Sollazzo M (2004) A novel ADP- and zinc-binding fold from function-directed in vitro evolution. Nat Struct Mol Biol 11: 382–383.
46. Chothia C, Gough J (2009) Genomic and structural aspects of protein evolution. Biochem J 419: 15–28.
47. Finkelstein AV, Ptitsyn O (2002) Protein Physics: A Course of Lectures. Elsevier Science.
48. Finkelstein AV, Reva BA (1991) A search for the most stable folds of protein chains. Nature 351: 497–499.
49. Matsuura T, Ernst A, Zechel DL, Plückthun A (2004) Combinatorial Approaches To Novel Proteins. ChemBioChem 5: 177–182.
50. Urvoas A, Valerio-Lepiniec M, Minard P (2012) Artificial proteins from combinatorial approaches. Trends Biotechnol 30: 512–520.
51. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH (1993) Protein design by binary patterning of polar and nonpolar amino acids. Science 262: 1680–1685.
52. West MW, Hecht MH (1995) Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. Protein Sci 4: 2032–2039.
53. Matsuura T, Ernst A, Plückthun A (2002) Construction and characterization of protein libraries composed of secondary structure modules. Protein Sci 11: 2631–2643.
54. Ernst A (2006) Combinatorial approaches to the evolution of novel proteins. University of Zürich.
55. Smith BA, Hecht MH (2011) Novel proteins: from fold to function. Curr Opin Chem Biol 15: 421–426.
56. Wang W, Hecht MH (2002) Rationally designed mutations convert de novo amyloid-like fibrils into monomeric beta-sheet proteins. Proc Natl Acad Sci USA 99: 2760–2765.
57. Richardson JS, Richardson DC (2002) Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci USA 99: 2754–2759.
58. Roberts RW, Szostak JW (1997) RNA-peptide fusions for the in vitro selection of peptides and proteins. Proc Natl Acad Sci USA 94: 12297–12302.
59. Yumerefendi H, Desravines DC, Hart DJ (2011) Library-based methods for identification of soluble expression constructs. Methods 55: 38–43.
60. Kristensen P, Winter G (1998) Proteolytic selection for protein folding using filamentous bacteriophages. Fold Des 3: 321–328.
61. Sieber V, Plückthun A, Schmid FX (1998) Selecting proteins with improved stability by a phage-based method. Nat Biotechnol 16: 955–960.
62. Christ D, Winter G (2006) Identification of protein domains by shotgun proteolysis. J Mol Biol 358: 364–371.
63. Wunderlich M, Martin A, Schmid FX (2005) Stabilization of the Cold Shock Protein CspB from Bacillus subtilis by Evolutionary Optimization of Coulombic Interactions. J Mol Biol 347: 1063–1076.
64. Matsuura T, Plückthun A (2003) Selection based on the folding properties of proteins with ribosome display. FEBS Lett 539: 24–28.
65. Scalley-Kim M, Minard P, Baker D (2003) Low free energy cost of very long loop insertions in proteins. Protein Sci 12: 197–206.
66. Minard P, Scalley-Kim M, Watters A, Baker D (2001) A "loop entropy reduction" phage-display selection for folded amino acid sequences. Protein Sci 10: 129–134.
67. Hagihara Y, Kim PS (2002) Toward development of a screen to identify randomly encoded, foldable sequences. Proc Natl Acad Sci USA 99: 6619–6624.
68. Papanikou E, Karamanou S, Economou A (2007) Bacterial protein secretion through the translocase nanomachine. Nat Rev Micro 5: 839–851.
69. Kudva R, Denks K, Kuhn P, Vogt A, Müller M, Koch H-G (2013) Protein translocation across the inner membrane of Gram-negative bacteria: the Sec and Tat dependent protein transport pathways. Res Microbiol 164: 505–534.

70. Randall LL, Hardy SJS (2002) SecB, one small chaperone in the complex milieu of the cell. Cell Mol Life Sci 59: 1617–1623.
71. Altman E, Kumamoto CA, Emr SD (1991) Heat-shock proteins can substitute for SecB function during protein export in Escherichia coli. EMBO J 10: 239–245.
72. Fröbel J, Rose P, Müller M (2012) Twin-arginine-dependent translocation of folded proteins. Philos Trans R Soc London 367: 1029–1046.
73. Berks BC (1996) A common export pathway for proteins binding complex redox cofactors? Mol Microbiol 22: 393–404.
74. Sargent F, Bogsch EG, Stanley NR, Wexler M, Robinson C, Berks BC, Palmer T (1998) Overlapping functions of components of a bacterial Sec-independent protein export pathway. EMBO J 17: 3640–3650.
75. Lüke I, Handford JI, Palmer T, Sargent F (2009) Proteolytic processing of Escherichia coli twin-arginine signal peptides by LepB. Arch Microbiol 191: 919–925.
76. Yahr TL, Wickner WT (2001) Functional reconstitution of bacterial Tat translocation in vitro. EMBO J 20: 2472–2479.
77. Cristóbal S, de Gier J, Nielsen H, Heijne von G (1999) Competition between Sec- and TAT-dependent protein translocation in Escherichia coli. EMBO J 18: 2982–2990.
78. Ize B, Gérard F, Wu L-F (2002) In vivo assessment of the Tat signal peptide specificity in Escherichia coli. Arch Microbiol 178: 548–553.
79. Tullman-Ercek D, DeLisa MP, Kawarasaki Y, Iranpour P, Ribnicky B, Palmer T, Georgiou G (2007) Export pathway selectivity of Escherichia coli twin arginine translocation signal peptides. J Biol Chem 282: 8309–8316.
80. Blaudeck N, Kreutzenbeck P, Freudl R, Sprenger GA (2003) Genetic Analysis of Pathway Specificity during Posttranslational Protein Translocation across the Escherichia coli Plasma Membrane. J Bacteriol 185: 2811–2819.
81. Hinsley AP, Stanley NR, Palmer T, Berks BC (2001) A naturally occurring bacterial Tat signal peptide lacking one of the “invariant” arginine residues of the consensus targeting motif. FEBS Lett 497: 45–49.
82. Ignatova Z, Hörnle C, Nurk A, Kasche V (2002) Unusual signal peptide directs penicillin amidase from Escherichia coli to the Tat translocation machinery. Biochem Biophys Res Commun 291: 146–149.
83. Widdick DA, Eijlander RT, van Dijl JM, Kuipers OP, Palmer T (2008) A Facile Reporter System for the Experimental Identification of Twin-Arginine Translocation (Tat) Signal Peptides from All Kingdoms of Life. J Mol Biol 375: 595–603.
84. Hatzixanthis K, Palmer T, Sargent F (2003) A subset of bacterial inner membrane proteins integrated by the twin-arginine translocase. Mol Microbiol 49: 1377–1390.
85. DeLisa MP, Tullman D, Georgiou G (2003) Folding quality control in the export of proteins by the bacterial twin-arginine translocation pathway. Proc Natl Acad Sci USA 100: 6115–6120.
86. Panahandeh S, Maurer C, Moser M, DeLisa MP, Müller M (2008) Following the path of a twin-arginine precursor along the TatABC translocase of Escherichia coli. J Biol Chem 283: 33267–33275.
87. Sanders C, Wethkamp N, Lill H (2001) Transport of cytochrome c derivatives by the bacterial Tat protein translocation system. Mol Microbiol 41: 241–246.
88. Santini CL, Ize B, Chanal A, Müller M, Giordano G, Wu L-F (1998) A novel sec-independent periplasmic protein translocation pathway in Escherichia coli. EMBO J 17: 101–112.
89. Rodrigue A, Chanal A, Beck K, Müller M, Wu L-F (1999) Co-translocation of a periplasmic enzyme complex by a hitchhiker mechanism through the bacterial tat pathway. J Biol Chem 274: 13223–13228.
90. Waraho D, DeLisa MP (2009) Versatile selection technology for intracellular protein-protein interactions mediated by a unique bacterial hitchhiker transport mechanism. Proc Natl Acad Sci USA 106: 3692–3697.
91. Speck J, Räuber C, Kükenshöner T, Niemöller C, Mueller KJ, Schleberger P, Dondapati P, Hecky J, Arndt KM, Müller KM (2013) TAT hitchhiker selection expanded to folding helpers, multimeric interactions and combinations with protein fragment complementation. Protein Eng Des Sel 26: 225–242.
92. Baglieri J, Beck D, Vasisht N, Smith CJ, Robinson C (2012) Structure of TatA Paralog, TatE, Suggests a Structurally Homogeneous Form of Tat Protein Translocase That Transports Folded Proteins of Differing Diameter. J Biol Chem 287: 7335–7344.
93. Hu Y, Zhao E, Li H, Xia B, Jin C (2010) Solution NMR Structure of the TatA Component of the Twin-Arginine Protein Transport System from Gram-Positive Bacterium Bacillus subtilis. J Am Chem Soc 132: 15942–15944.
94. Zhang Y, Wang L, Hu Y, Jin C (2014) Solution structure of the TatB component of the twin-arginine translocation system. Biochim Biophys Acta 1838: 1881–1888.
95. Rollauer SE, Tarry MJ, Graham JE, Jaaskelainen M, Jäger F, Johnson S, Krehenbrink M, Liu S-M, Lukey MJ, Marcoux J, McDowell MA, Rodriguez F, Roversi P, Stansfeld PJ, Robinson CV, Sansom MSP, Palmer T, Hogbom M, Berks BC, Lea SM (2012) Structure of the TatC core of the twin-arginine protein transport system. Nature 492: 210–214.
96. Patel R, Smith SM, Robinson C (2014) Protein transport by the bacterial Tat pathway. Biochim Biophys Acta.
97. Dubini A, Sargent F (2003) Assembly of Tat-dependent [NiFe] hydrogenases: identification of precursor-binding accessory proteins. FEBS Lett 549: 141–146.
98. Genest O, Seduk F, Ilbert M, Méjean V, Iobbi-Nivol C (2006) Signal peptide protection by specific chaperone. Biochem Biophys Res Commun 339: 991–995.

99. Grahl S, Maillard J, Spronk CAEM, Vuister GW, Sargent F (2012) Overlapping transport and chaperone-binding functions within a bacterial twin-arginine signal peptide. Mol Microbiol 83: 1254–1267.
100. Jack RL, Buchanan G, Dubini A, Hatzixanthis K, Palmer T, Sargent F (2004) Coordinating assembly and export of complex bacterial proteins. EMBO J 23: 3962–3972.
101. Holzapfel E, Moser M, Schiltz E, Ueda T, Betton J-M, Müller M (2009) Twin-Arginine-Dependent Translocation of SufI in the Absence of Cytosolic Helper Proteins. Biochemistry 48: 5096–5105.
102. Rocco MA, Waraho-Zhmayev D, DeLisa MP (2012) Twin-arginine translocase mutations that suppress folding quality control and permit export of misfolded substrate proteins. Proc Natl Acad Sci USA 109: 13392–13397.
103. Richter S, Brüser T (2005) Targeting of Unfolded PhoA to the TAT Translocon of Escherichia coli. J Biol Chem 280: 42723–42730.
104. Matos CFRO, Robinson C, Di Cola A (2008) The Tat system proofreads FeS protein substrates and directly initiates the disposal of rejected molecules. EMBO J 27: 2055–2063.
105. Brüser T, Sanders C (2003) An alternative model of the twin arginine translocation system. Microbiol Res 158: 7–17.
106. Richter S, Lindenstrauss U, Lucke C, Bayliss R, Brüser T (2007) Functional Tat transport of unstructured, small, hydrophilic proteins. J Biol Chem 282: 33257–33264.
107. Fisher AC, Kim W, DeLisa MP (2006) Genetic selection for protein solubility enabled by the folding quality control feature of the twin-arginine translocation pathway. Protein Sci 15: 449–458.
108. Lee PA, Tullman-Ercek D, Georgiou G (2006) The bacterial twin-arginine translocation pathway. Annu Rev Microbiol 60: 373–395.
109. Barrett CML, Ray N, Thomas JD, Robinson C, Bolhuis A (2003) Quantitative export of a reporter protein, GFP, by the twin-arginine translocation pathway in Escherichia coli. Biochem Biophys Res Commun 304: 279–284.
110. Bowden GA, Baneyx F, Georgiou G (1992) Abnormal fractionation of beta-lactamase in Escherichia coli: evidence for an interaction with the inner membrane in the absence of a leader peptide. J Bacteriol 174: 3407–3410.
111. Ormö M, Cubitt AB, Kallio K, Gross LA, Tsien RY, Remington SJ (1996) Crystal structure of the Aequorea victoria green fluorescent protein. Science 273: 1392–1395.
112. Feilmeier BJ, Iseminger G, Schroeder D, Webber H, Phillips GJ (2000) Green fluorescent protein functions as a reporter for protein localization in Escherichia coli. J Bacteriol 182: 4068–4076.
113. Thomas JD, Daniel RA, Errington J, Robinson C (2001) Export of active green fluorescent protein to the periplasm by the twin-arginine translocase (Tat) pathway in Escherichia coli. Mol Microbiol 39: 47–53.
114. Santini CL, Bernadac A, Zhang M, Chanal A, Ize B, Blanco C, Wu L-F (2001) Translocation of jellyfish green fluorescent protein via the Tat system of Escherichia coli and change of its periplasmic localization in response to osmotic up-shock. J Biol Chem 276: 8159–8164.
115. Østergaard H, Henriksen A, Hansen FG, Winther JR (2001) Shedding light on disulfide bond formation: engineering a redox switch in green fluorescent protein. EMBO J 20: 5853–5862.
116. Hanson GT, Aggeler R, Oglesbee D, Cannon M, Capaldi RA, Tsien RY, Remington SJ (2004) Investigating mitochondrial redox potential with redox-sensitive green fluorescent protein indicators. J Biol Chem 279: 13044–13053.
117. Pedelacq J-D, Cabantous S, Tran T, Terwilliger TC, Waldo GS (2006) Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol 24: 79–88.
118. DeLisa MP, Samuelson P, Palmer T, Georgiou G (2002) Genetic analysis of the twin arginine translocator secretion pathway in bacteria. J Biol Chem 277: 29825–29831.
119. Sauer RT, Baker TA (2011) AAA+ proteases: ATP-fueled machines of protein destruction. Annu Rev Biochem 80: 587–612.
120. Baker TA, Sauer RT (2012) ClpXP, an ATP-powered unfolding and protein-degradation machine. Biochim Biophys Acta 1823: 15–28.
121. Wah DA, Levchenko I, Baker TA, Sauer RT (2002) Characterization of a Specificity Factor for an AAA+ ATPase: Assembly of SspB Dimers with ssrA-Tagged Proteins and the ClpX Hexamer. Chem Biol 9: 1237–1245.
122. Hersch GL, Baker TA, Sauer RT (2004) SspB delivery of substrates for ClpXP proteolysis probed by the design of improved degradation tags. Proc Natl Acad Sci USA 101: 12136–12141.
123. Flynn JM, Levchenko I, Seidel M, Wickner SH, Sauer RT, Baker TA (2001) Overlapping recognition determinants within the ssrA degradation tag allow modulation of proteolysis. Proc Natl Acad Sci USA 98: 10584–10589.
124. Steiner D, Forrer P, Stumpp MT, Plückthun A (2006) Signal sequences directing cotranslational translocation expand the range of proteins amenable to phage display. Nat Biotechnol 24: 823–831.
125. Knappik A, Ge L, Honegger A, Pack P, Fischer M, Wellnhofer G, Hoess A, Wölle J, Plückthun A, Virnekäs B (2000) Fully synthetic human combinatorial antibody libraries (HuCAL) based on modular consensus frameworks and CDRs randomized with trinucleotides. J Mol Biol 296: 57–86.
126. Randall LL, Topping TB, Suciu D, Hardy SJ (1998) Calorimetric analyses of the interaction between SecB and its ligands. Protein Sci 7: 1195–1200.
127. Pugsley AP (1993) The complete general secretory pathway in gram-negative bacteria. Microbiol Rev 57: 50–108.

128. Kononova SV, Khokhlova OV, Zolov SN, Nesmeyanova MA (2001) Effect of export-specific cytoplasmic chaperone, protein SecB, on secretion of Escherichia coli alkaline phosphatase. Biochemistry (Mosc) 66: 803–807.
129. Diao L, Dong Q, Xu Z, Yang S, Zhou J, Freudl R (2012) Functional implementation of the posttranslational SecB-SecA protein-targeting pathway in Bacillus subtilis. Appl Environ Microbiol 78: 651–659.
130. Schierle CF, Berkmen M, Huber D, Kumamoto C, Boyd D, Beckwith J (2003) The DsbA signal sequence directs efficient, cotranslational export of passenger proteins to the Escherichia coli periplasm via the signal recognition particle pathway. J Bacteriol 185: 5706–5713.
131. Fröbel J, Rose P, Lausberg F, Blümmel A-S, Freudl R, Müller M (2012) Transmembrane insertion of twin-arginine signal peptides is driven by TatC and regulated by TatB. Nat Commun 3: 1311.
132. Kohl A, Binz HK, Forrer P, Stumpp MT, Plückthun A, Grütter MG (2003) Designed to be stable: crystal structure of a consensus ankyrin repeat protein. Proc Natl Acad Sci USA 100: 1700–1705.
133. Binz HK, Stumpp MT, Forrer P, Amstutz P, Plückthun A (2003) Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins. J Mol Biol 332: 489–503.
134. Parmeggiani F (2008) Design of armadillo repeat protein scaffolds. University of Zürich.
135. Parmeggiani F, Pellarin R, Larsen AP, Varadamsetty G, Stumpp MT, Zerbe O, Caflisch A, Plückthun A (2008) Designed Armadillo Repeat Proteins as General Peptide-Binding Scaffolds: Consensus Design and Computational Optimization of the Hydrophobic Core. J Mol Biol 376: 1282–1304.
136. Lutz S, Fast W, Benkovic SJ (2002) A universal, vector-based system for nucleic acid reading-frame selection. Protein Eng 15: 1025–1030.
137. Gerth ML, Patrick WM, Lutz S (2004) A second-generation system for unbiased reading frame selection. Protein Eng Des Sel 17: 595–602.
138. Heim R, Cubitt AB, Tsien RY (1995) Improved green fluorescence. Nature 373: 663–664.
139. Aronson DE, Costantini LM, Snapp EL (2011) Superfolder GFP Is Fluorescent in Oxidizing Environments When Targeted via the Sec Translocon. Traffic 12: 543–548.
140. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2: 2006.0008.
141. Keiler KC, Waller PR, Sauer RT (1996) Role of a peptide tagging system in degradation of proteins synthesized from damaged messenger RNA. Science 271: 990–993.
142. Waldo GS, Standish BM, Berendzen J, Terwilliger TC (1999) Rapid protein-folding assay using green fluorescent protein. Nat Biotechnol 17: 691–695.
143. Pedelacq J-D, Piltch E, Liong EC, Berendzen J, Kim C-Y, Rho B-S, Park MS, Terwilliger TC, Waldo GS (2002) Engineering soluble proteins for structural genomics. Nat Biotechnol 20: 927–932.
144. Graubner W, Schierhorn A, Brüser T (2007) DnaK plays a pivotal role in Tat targeting of CueO and functions beside SlyD as a general Tat signal binding chaperone. J Biol Chem 282: 7116–7124.
145. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637.
146. Andersen CAF, Palmer AG, Brunak S, Rost B (2002) Continuum Secondary Structure Captures Protein Flexibility. Structure 10: 175–184.
147. Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev SK, Dryden DTF, Dybvig K, Firman K, Gromova ES, Gumport RI, Halford SE, Hattman S, Heitman J, Hornby DP, Janulaitis A, Jeltsch A, Josephsen J, Kiss A, Klaenhammer TR, Kobayashi I, Kong H, Krüger DH, et al. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. Nucleic Acids Res 31: 1805–1812.
148. Roberts RJ, Vincze T, Posfai J, Macelis D (2010) REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res 38: D234–D236.
149. Prijambada ID, Yomo T, Tanaka F, Kawama T, Yamamoto K, Hasegawa A, Shima Y, Negoro S, Urabe I (1996) Solubility of artificial proteins with random sequences. FEBS Lett 382: 21–25.
150. Cho G, Keefe AD, Liu R, Wilson DS, Szostak JW (2000) Constructing high complexity synthetic libraries of long ORFs using In Vitro selection. J Mol Biol 297: 309–319.
151. Chiarabelli C, Vrijbloed JW, De Lucrezia D, Thomas RM, Stano P, Polticelli F, Ottone T, Papa E, Luisi PL (2006) Investigation of de novo Totally Random Biosequences, Part II. Chem Biodivers 3: 840–859.
152. Tanaka J, Doi N, Takashima H, Yanagawa H (2010) Comparative characterization of random-sequence proteins consisting of 5, 12, and 20 kinds of amino acids. Protein Sci 19: 786–795.
153. Cordes MHJ, Davidson AR, Sauer RT (1996) Sequence space, folding and protein design. Curr Opin Struct Biol 6: 3–10.
154. Wells JA (1990) Additivity of mutational effects in proteins. Biochemistry 29: 8509–8517.
155. Davidson AR, Lumb KJ, Sauer RT (1995) Cooperatively folded proteins in random sequence libraries. Nat Struct Biol 2: 856–864.
156. Doi N, Kakukawa K, Oishi Y, Yanagawa H (2005) High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. Protein Eng Des Sel 18: 279–284.

157. Hecht MH, Das A, Go A, Bradley LH, Wei Y (2004) De novo proteins from designed combinatorial libraries. Protein Sci 13: 1711–1723.
158. de Bono S, Riechmann L, Girard E, Williams RL, Winter G (2005) A segment of cold shock protein directs the folding of a combinatorial protein. Proc Natl Acad Sci USA 102: 1396–1401.
159. Graziano JJ, Liu W, Perera R, Geierstanger BH, Lesley SA, Schultz PG (2008) Selecting folded proteins from a library of secondary structural elements. J Am Chem Soc 130: 176–185.
160. Labean TH, Kauffman SA (1993) Design of synthetic gene libraries encoding random sequence proteins with desired ensemble characteristics. Protein Sci 2: 1249–1254.
161. Riechmann L, Winter G (2000) Novel folded protein domains generated by combinatorial shuffling of polypeptide segments. Proc Natl Acad Sci USA 97: 10068–10073.
162. Tabuchi I, Soramoto S, Ueno S, Husimi Y (2004) Multi-line split DNA synthesis: a novel combinatorial method to make high quality peptide libraries. BMC Biotechnol 4: 19.
163. Watters AL, Baker D (2004) Searching for folded proteins in vitro and in silico. Eur J Biochem 271: 1615–1622.
164. Virnekäs B, Ge L, Plückthun A, Schneider KC, Wellnhofer G, Moroney SE (1994) Trinucleotide phosphoramidites: ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis. Nucleic Acids Res 22: 5600–5607.
165. Braunagel M, Little M (1997) Construction of a semisynthetic antibody library using trinucleotide oligos. Nucleic Acids Res 25: 4690–4691.
166. Denault M, Pelletier J (2007) Protein Library Design and Screening. In: Arndt K, Müller K, editors. Methods in Molecular Biology. Protein Engineering Protocols. New Jersey: Humana Press, Vol. 352. pp. 127–154.
167. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) Protein identification and analysis tools in the ExPASy server. In: Walker JM, editor. Methods in Molecular Biology. Humana Press. pp. 571—607.
168. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet 16: 276–277.
169. Kawe M, Horn U, Plückthun A (2009) Facile promoter deletion in Escherichia coli in response to leaky expression of very robust and benign proteins from common expression vectors. Microb Cell Fact 8: 8.
170. Dyrløv Bendtsen J, Nielsen H, Heijne von G, Brunak S (2004) Improved Prediction of Signal Peptides: SignalP 3.0. J Mol Biol 340: 783–795.
171. Petersen TN, Brunak S, Heijne von G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8: 785–786.
172. Solovyev V, Salamov A (n.d.) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In: Li RW, editor. Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies. Nova Science Publishers. pp. 61–78.
173. Taylor WR (1997) Residual colours: a proposal for aminochromography. Protein Eng 10: 743–746.
174. Steiner D (2008) Filamentous phage display of designed ankyrin repeat proteins: from conception to applications. University of Zürich.
175. Matos CFRO, Branston SD, Albiniak A, Dhanoya A, Freedman RB, Keshavarz-Moore E, Robinson C (2012) High-yield export of a native heterologous protein to the periplasm by the tat translocation pathway in Escherichia coli. Biotechnol Bioeng 109: 2533–2542.
176. Branston SD, Matos CFRO, Freedman RB, Robinson C, Keshavarz-Moore E (2012) Investigation of the impact of Tat export pathway enhancement on E. coli culture, protein production and early stage recovery. Biotechnol Bioeng 109: 983–991.
177. Whitaker N, Bageshwar UK, Musser SM (2013) Effect of cargo size and shape on the transport efficiency of the bacterial Tat translocase. FEBS Lett 587: 912–916.
178. Bageshwar UK, Musser SM (2007) Two electrical potential-dependent steps are required for transport by the Escherichia coli Tat machinery. J Cell Biol 179: 87–99.
179. Sambrook J, Russell DW (2001) Molecular Cloning: A Laboratory Manual. 3rd ed. Cold Spring Harbor Laboratory Press.
180. Chen Y-C, Chen L-A, Chen S-J, Chang M-C, Chen T-L (2009) A modified osmotic shock for periplasmic release of a recombinant creatinase from Escherichia coli. Biochem Eng J 19: 211–215.
181. Nossal NG, Heppel LA (1966) The release of enzymes by osmotic shock from Escherichia coli in exponential phase. J Biol Chem 241: 3055–3062.

6 Appendix

6.1 Supplemental tables and figures

Table 6.1: Oligonucleotides used in this work.

cham5r	CTAATTAAGCTTAGGATCCTTGGTCTCTACC
fw_Bmod	CTAACTGAGTAACGTCAGGTGACCGGGCCCTCTGGTCACCCAG
fw_BpiNNK	AGCGAGCCCCAAGACTTATTTCTTCTGTGTGCGAAGACG
fw_Gmod	GGGCCCTCTGGTAGCAAAGGAGAAGAACTTTTC
fw_joinSSL	TAGTGACTGATAGCTAGTGACGATAGTGGTTCC
fw_NcoI_SSLjoin	TAGTGACTGATAGCTAGTGACGATAGTGGTTCCATGGAAGACACTGGT
fw_NNK_BsaI	GCGGCGCAAGCGGACTACAAAGATGGATCCGGGTCTC
fw_NNK_t	GCGGCgcaagcgGACTACAAAGATGGATCCGGtTCTC
fw_SSL_BclI	GGCTGATCAVTHVTHVTHVTHVTHGGTTCCGTGGAAGACACTGGT
fw_SSL_noNco_TorA	GCGCAAGCGGACTACAAAGATGG
fw_Tmod	CGATCTCTTTCAGGCATCACGTCG
rv_Bmod	GCGCAACGTTGTTGCCATTG
rv_BpiNNK	CACTAAATGTTTACGCTAACACGGCAGGCGAGAAGACGG
rv_Gmod	CGTTTCATGTGATCCGGATAACGG
rv_HindSacSsrA	GTCAAGCTTGAGCTCAGTTAGTCATTAGGCGGCC
rv_NNK_BsaI	GTTCTTCTCCTTTGCTACCAGAGGGCCCAGGGTCTCC
rv_phiNK	GGTCACCTGACGTTACTCAGTTAGAAGACCG
rv_SpSPCRlib1	GTTCTTCTCCTTTGCTACCAGAGGGCCCGGACCCTTG
rv_Tmod	GGATCCATCTTTGTAGTCCGCTTGC

Oligonucleotide names are given in the left column, their 5' to 3' sequences in the right column.
See also **Table 2.3**.

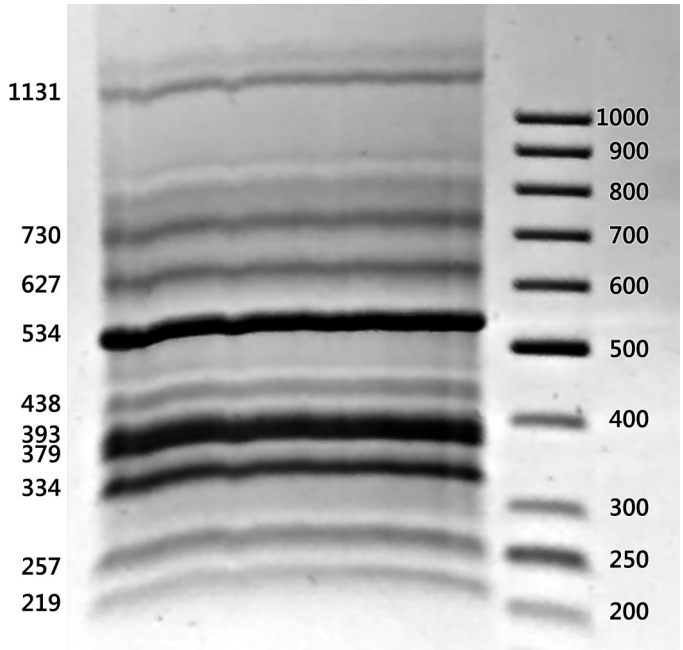


Figure 6.1: Ethidium-bromide stained agarose gel of first NNK module ligation.
Ligation products of non-directed ligation of 253 bp Tmod-1N-BpiI (210 bp + 4 nt and 39 bp + 4 nt after BpiI digest) and 370 bp BpiI-1N-Gmod (326 bp + 4 nt and 40 bp + 4 nt after BpiI digest). The calculated size of the desired ligation product Tmod-2N-Gmod is 540 bp, and the two self-ligated side-products of 424 bp and 656 bp. Compare to **Figure 2.27e,f**.

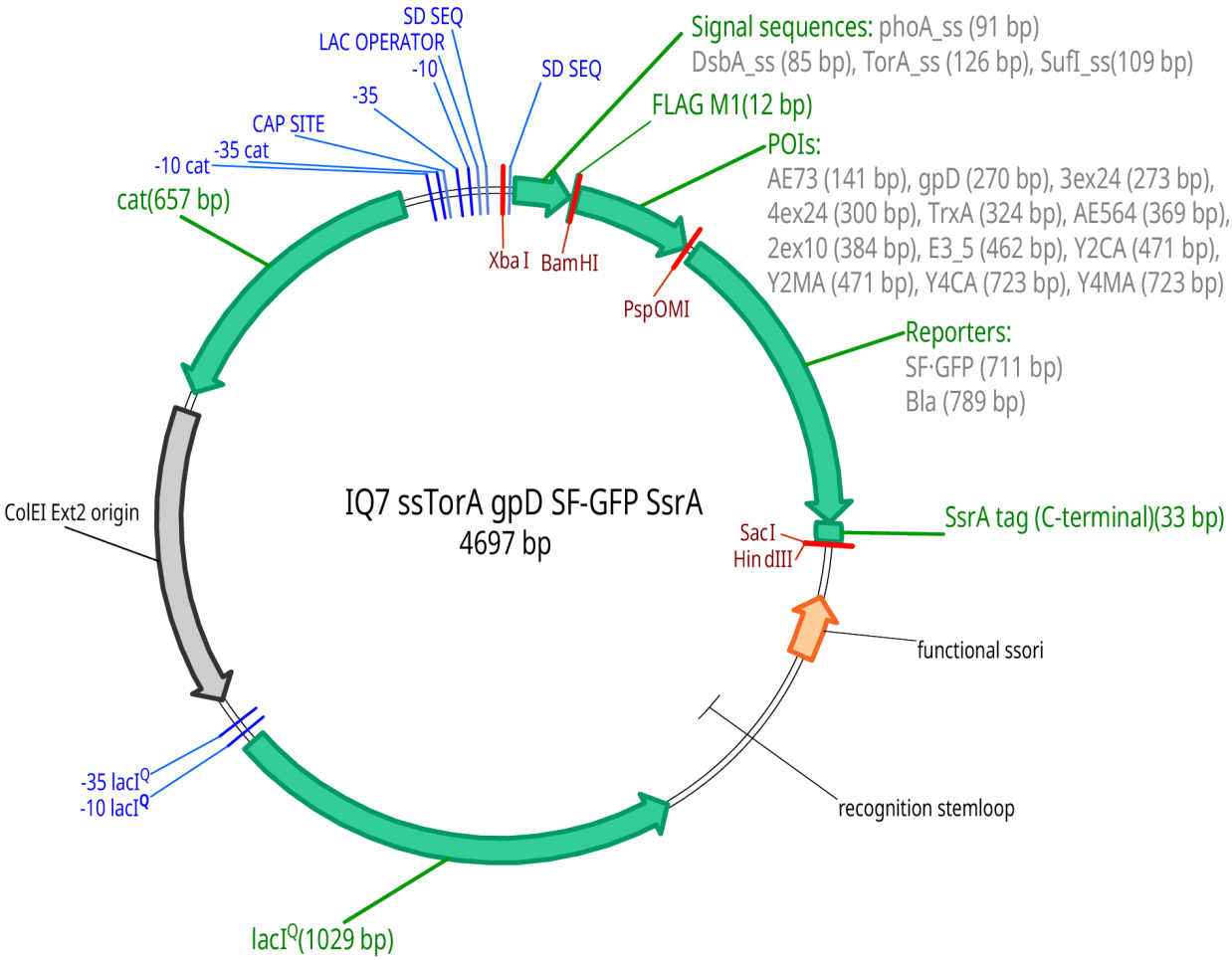


Figure 6.2: Vector map of the main constructs used for characterization of Tat-dependent translocation. The signal sequences target translocation to the periplasm of *E. coli.* via different pathways (2.1.2). The proteins of interest (POIs) possess a variety of folding characteristics (2.1.3), and the reporter proteins allow quantification of translocated proteins (2.1.4).

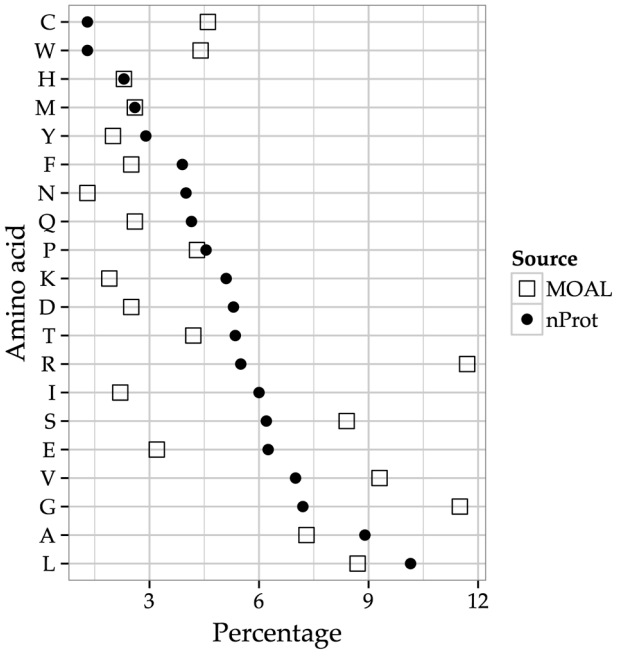


Figure 6.3: Amino acid composition encoded by the MOAL compared to natural proteins.
nProt: amino acid composition of natural protein, calculated as arithmetic mean of the composition of all SwissProt proteins and the translated codons of 5045 coding sequences of *Escherichia coli* K12 (see **Figure 6.4** below, which is a copy of **Figure 2.26**).

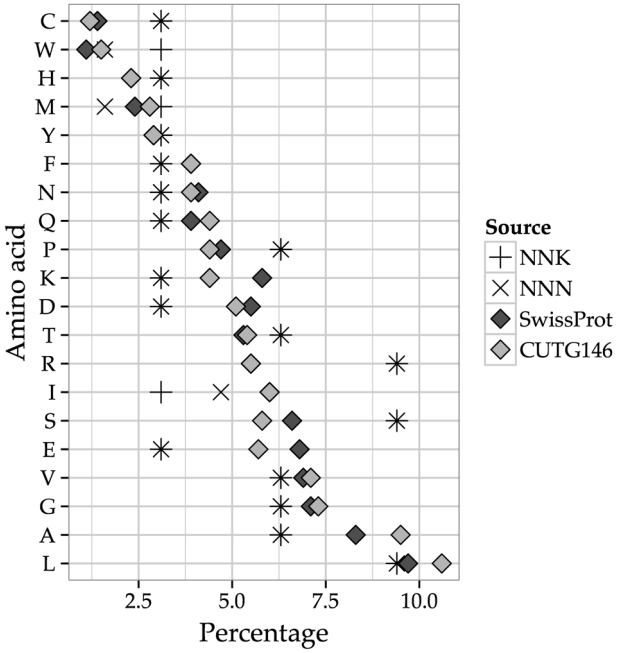


Figure 6.4 (copy of Figure 2.26): Expected amino acid distributions for NNN or NNK codons compared to natural proteins.
SwissProt: Amino acid composition (%) in the UniProtKB/Swiss-Prot data bank as found in the release notes for UniProtKB/Swiss-Prot release 2013_04 - April 2013 [167]. CUTG146: Translated codons of 5045 coding sequences of *Escherichia coli* K12 (Division: gbbct, Release: CUTG146) as found in e.g. EMBOSS [168].

6.2 Abbreviations

2YT	double strength yeast extract trypton medium
a.k.a.	also known as
aa	amino acid
Amp	ampicillin
Bla	β-lactamase
cfu	colony forming units
$Cm^{(R)}$	chloramphenicol (resistance)
EDTA	ethylenediaminetetraacetic acid
FACS	fluorescence activated cell sorting
GFP	green fluorescent protein
IPTG	isopropyl-β-D-thiogalactopyranoside
kDa	kilo-Dalton
LB	Luria Broth
MBP	maltose binding protein
MOAL	mother of all libraries, a fully random library of 303 bp (see 2.3)
OD_{600}	optical density at 600 nm
ORF	open reading frame
PDB	Protein Data Bank
Φ (phi)	a short library module encoding five consecutive hydrophobic amino acids (2.2.2)
POI	protein of interest
RMSD	root mean square deviation
SF·GFP	superfolder GFP
SRP	signal recognition particle
SSL	secondary structure library
Tat	twin arginine translocation
wtGFP	wild-type GFP (see also 2.1.10 for S65T·GFP)

6.3 Acknowledgements

I want to sincerely thank Andreas Plückthun for giving me the opportunity to work on such an interesting and challenging project in fundamental protein science. Nowadays, it seems almost like a luxury to be able to work in basic research, especially in fields where seemingly a broad knowledge has been accumulated. I hope that many more will be able to do basic research and demonstrate that a better understanding of fundamental principles will be beneficial in the long run, even if no immediate applicability is given.

Many thanks go also to the members of my thesis committee, Elke Deuerling and Ben Schuler for their valuable feedback and encouragement.

I would further like to acknowledge all the people who keep the infrastructure running, of our group, our institute, and external facilities I came in contact with, such as the FACS facility. They are all doing a great job and there would be a lot more friction in lab life without people like Peter Lindner, Petra Vogt, Birgit Dreier, the IT staff, workshop staff, and all the others.

Thanks go to Andreas Ernst, who built the version 2.1 of the Secondary Structure Library that I took over and used in this thesis and who introduced me to the experimental part of library constructions and selections.

I wish to also thank all my previous and current colleagues in the Plückthun lab for creating such a nice working atmosphere, and especially my present and former lab-mates for the good times in M94 (Marc-Simon, Jakob, Karola, Myriam, Igor, and Thomas). A special “thank you” goes to Yvonne for her help and for being a good friend.

Finally, I would like to deeply thank Sabine and my family for their support and their trust in me.

6.4 Curriculum Vitae

— Personal data —

Full Name (Family name, given name):	SCHMITZ, Mark Alexander
Date of birth:	1977-09-06
Place of birth:	Timisoara, Romania
Citizenship:	Germany

— Education —

2006 - 2014 Dissertation in the group of Prof. Dr. Andreas Plückthun, Department of Biochemistry, University of Zürich
Title: "Seeking truly novel proteins in a fully random library by bacterial Tat-dependent selections for folding"
MLS Ph.D. program, Life Science Zurich Graduate School

2004 - 2005 Diploma thesis in the group of Dr. Kristian Müller, Faculty of Biology III, Albert Ludwigs University of Freiburg
Title: "Generierung und Charakterisierung von Fusionsproteinen bestehend aus der Nuklease NucA und Fluorescein bindendem scFv Antikörper"

1999 - 2005 Studies of Biology (Genetics, Biochemistry and Cell Biology) at the Albert Ludwigs University of Freiburg

1998 - 1999 Studies of Mathematics and Chemistry at the Albert Ludwigs University of Freiburg

1997 Abitur at the Albert-Schweitzer-Gymnasium, Leonberg